January 2012

Basal Graph Structures for Geometry Based Organization of Wide-Baseline Image Collections

Aveek Shankar Brahmachari
University of South Florida, abrahmac@mail.usf.edu

Basal Graphs Structures for Geometry Based Organization of Wide-Baseline Image Collections

by

Aveek Shankar Brahmachari

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Computer Science & Engineering
College of Engineering
University of South Florida

Major Professor: Sudeep Sarkar, Ph.D.
Rangachar Kasturi, Ph.D.
Dmitry Goldgof, Ph.D.
Thomas Sanocki, Ph.D.
Autar Kaw, Ph.D.

Date of Approval:
November 6, 2012

Keywords: Epipolar Geometry, Correspondence, Minimum Spanning Forest, GPS, Magnetometer

Dedication

Dedicated to my family and friends.

Acknowledgments

I would like to express my deep gratitude to Prof. Sudeep Sarkar for giving me an opportunity to work under his supervision. During my doctoral research, Prof. Sarkar encouraged me to span across widely spread set of related problems. This has lead to a truly challenging and satisfying experience. I also want to thank him for making me work on a core topic of wide interest in the field of computer vision.

I want to thank Prof. Kasturi, Prof. Goldgof, Prof. Sanocki and Prof. Kaw for agreeing to be my committee members and bearing with my indecisiveness on final doctoral defense dates. I would want to thank Prof. Chari for kindly agreeing to preside my doctoral defense. I am thankful to Prof. Sarkar, Prof. Kasturi, Prof. Goldgof, Prof. Hall and Prof. Iamnitchi for enriching my knowledge with few of the best courses I took during my doctoral research.

I am thankful to all my friends who made both my life and work seem easy. Special thanks to Abhishek Kumar and Makhan Virdi for offering help whenever needed. I am thankful to Ravi Kiran Krishnan, Rajmadhan Ekambaram, Ravi Panchumarthy and Kester Duncan for their critical suggestions, humor and support at work. I would like to thank the technical support team of the department for their help.

Finally, I want to thank my wife, my parents and my sister. Without their inspiration and support, I would not have been able to pursue my doctoral aspirations. Special thanks to my wife who had to sacrifice initial years of her professional career for choosing to be with me through thick and thin of my doctoral journey.

Table of Contents

List of Tables

List of Figures

Abstract

We propose algorithms for organization of images in wide-area sparse-view datasets. In such datasets, if the images overlap in scene content, they are related by wide-baseline geometric transformations. The challenge is to identify these relations even if the images sparingly overlap in their content. The images in a dataset are then grouped into sets of related images with the relations captured in each set as a basal (minimal and foundational) graph structures. Images form the vertices in the graph structure and the edges define the geometric relations between the images. We use these basal graphs for geometric walk-throughs and detection of noisy location (GPS) and orientation (magnetometer) information that may be stored with each image.

We have five algorithmic contributions. First, we propose an algorithm BLOGS (Balanced Local and Global Search) that uses a novel hybrid Markov Chain Monte Carlo (MCMC) strategy called 'hop-diffusion' for epipolar geometry estimation between a pair of wide-baseline images that is 10 times faster and more accurate than the state-of-the-art. Hops are global searches and diffusions are local searches. BLOGS is able to handle very wide-baseline views characteristic of wide-area sparse-view datasets. It also produces a geometric match score between an image pair. Second, we propose a photometric match score, the Cumulative Correspondence Score (CCS). The proposed photometric scores are fast approximations of the computationally expensive geometric scores. Third, we use the photometric scores and the geometric scores to find groups of related images and to organize them in the form of basal graph structures using a novel hybrid algorithm we call the COnnected component DIscovery by Minimally Specifying an Expensive Graph (CODIMSEG). The objective of the algorithm is to minimize the number of geometric estimations and yield results similar to what would be achieved if all-pair geometric matching were

done. We compared the performances of the CCS and CODIMSEG algorithms with GIST (means summary of an image) and k-Nearest Neighbor (k-NN) based approaches. We found that CCS and CODIMSEG perform significantly better than GIST and k-NN respectively in identifying visually connected images. Our algorithm achieved more than 95% true positive rate at 0% false positive rate. Fourth, we propose a basal tree graph expansion algorithm to make the basal graphs denser for applications like geometric walk-throughs using the minimum Hamiltonian path algorithm and detection of noisy position (GPS) and orientation (magnetometer) tags. We propose two versions of geometric walkthroughs, one using minimum spanning tree based approximation of the minimum Hamiltonian path on the basal tree graphs and other using the Lin-Kernighan heuristic approximation on the expanded basal graph. Conversion of a non-linear tree structure to a linear path structure leads to discontinuities in path. The Lin-Kernighan algorithm on the expanded basal graphs is shown to be a better approach. Fifth, we propose a vision based geometric voting algorithm to detect noisy GPS and magnetometer tags using the basal graphs. This problem has never been addressed before to the best of our knowledge.

We performed our experiments on the Nokia dataset (which has 243 images in the 'Lausanne' dataset and 105 images in the 'Demoset'), ArtQuad dataset (6514 images) and Oxford dataset (5063 images). All the three datasets are very different. Nokia dataset is a very wide-baseline sparse-view dataset. ArtQuad dataset is a wide-baseline dataset with denser views compared to the Nokia dataset. Both these datasets have GPS tagged images. Nokia dataset has magnetometer tags too. ArtQuad dataset has 348 images with the commercial GPS information as well as high precision differential GPS data which serves as ground truth for our noisy tag detection algorithm. Oxford dataset is a wide-baseline dataset with plenty of distracters that test the algorithm's capability to group images correctly. The larger datasets test the scalability of our algorithms. Visually inspected feature matches and image matches were used as ground truth in our experiments. All the experiments were done on a single PC.

Chapter 1 Introduction

Large collections of images are ubiquitous today due to the advancement in technology that has led to the advent of easily affordable modern digital cameras and large digital storage spaces. Nowadays, thousands of images can easily be stored in multimedia devices like cell-phones equipped with a camera. However, if these images are not organized, it is difficult to interpret the overall information in the collection of the images. In contrast, if the images are organized, browsing through them might tell a story.

Figure 1.1: An unorganized collection of images. Note that we tend to organize them in our mind.

Organization of an image collection basically involves defining the relative arrangements of the images in them. An image collection is unorganized if no such arrangement is available. Figure 1.1 shows an example of an unorganized image collection. Notice that it takes a while in our mind to arrange the images in the collection and to identify the sets of images that overlap in scene content.

In the state-of-the-art, the easiest way of organizing large image collections is by associating the images with tags containing the name of location and the time of capture. Many cameras are also equipped with a GPS sensor and a full or partial INS sensor which sense the position and orientation of the cameras respectively. Thus, position and orientation are among other meta-data information that are often tagged with images these days. Tags are useful in making images accessible by search and thus they help in leveraging the information in relevant images. Organizing images by their visual content using tags is more meaningful for some applications than simple image arrangement using tags. To a certain extent, position and orientation tags can also help in organizing images according to their visual content by pre-filtering out relationships between those images that cannot possibly match. However, tags cannot be used just by themselves to organize images by their visual content. This is because we cannot be certain if the scene captured by a pair of cameras lie in the intersection of their field of views as illustrated in Figure 1.2. Also, unfortunately, not all images have tags and tags are often noisy. Thus, we look at two ensuing problems, one of organizing images without tags and the other of identifying noisy tags.

Vision can be used to organize images by visual content without using tags by matching image pairs in a collection. An organized collection of images is shown in Figure 1.3. Organized image collections can be exploited for many useful geometric applications like 3D reconstruction, 2D panorama stitching, geometric walk-throughs and detection of noisy tags.

One of the important factors that distinguish research works in vision based image organization is the baseline. Wide-baseline is a term referring to the large distance between

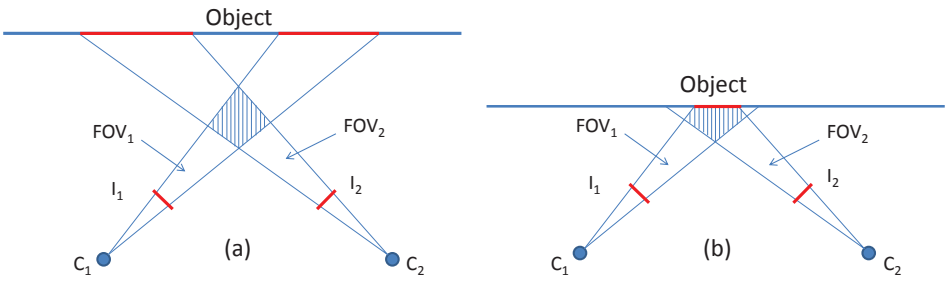


Figure 1.2: Figure showing the scene position and the intersection of the field of views of two cameras. (a) The scene is not in the intersecting field of view (b) The scene is in the intersecting field of view.

the optical centers of two cameras relative to the distance of the cameras from the scene they capture. Thus, the definition of wide-baseline also indicates having wide-angle between the cameras. Most of the state-of-the-art research works assume the availability of narrow-baseline images in their dataset. However, in an arbitrary image collection, not only are we likely to find few matching views, but we are also more likely to find wide-baseline views than narrow-baseline views. *Given the importance of the applications and the sparseness of the views in majority of the datasets collectible in practice, very little research has been done on image collections with sparse views spanning across large areas.* One of the reasons is that in general, for such datasets, dense reconstruction is not possible because of sparse views, and panorama stitching is not possible because of wide-baseline. However, such image collections might have regions where dense 3D reconstruction or 2D panorama stitching might be possible. Also, the information in such image collections can be leveraged for use in other important applications like geometric walk-throughs and detection of noisy position (GPS) and orientation (magnetometer) tags that are the focus problems of this dissertation.

At this point, the reader is referred to the glossary of terms in Appendix A for a list of terms used in the dissertation and their definition. The terms are mentioned mostly in the order in which they appear in the dissertation.

1.1 Goals and Objectives

The goal of this dissertation is to organize images to leverage geometric information in large wide-area sparse-view datasets using computer vision. Another goal is to bring large scale image organization within the reach of the computation power of a single PC or a future hand-held device in feasible time.

We seek to design algorithms that can efficiently match wide-baseline images sparsely spanning over a wide-area. Wide-baseline images are difficult to match, broadly because of two reasons: detection and description. Same features might not be detected in two images with overlapping scene content if they are separated by a wide-baseline. Even if the same features are detected, the corresponding feature descriptors might vary a lot due to large geometric transformation between the images, varying illumination and varying angle of reflectance. The matching problem might also be aggravated due to the presence of repeated patterns. Next is the problem of efficiently minimizing the use of computational resources without sacrificing the quality of result in multi-image organization problem. Finally, we seek basal (minimal and foundational) graph structures for geometry based applications like geometric walk-throughs, and detection of noisy position (GPS) and orientation (magnetometer) tags. Geometric walk-throughs help in informative visualization of an image collection, and identifying noisy position (GPS) and orientation (magnetometer) tags is useful for other applications dependent on correctness of these tags. The objectives of this dissertation are:

1. Our first objective is to design a computationally efficient algorithm to estimate epipolar geometry between wide-baseline images. In order to meet our first objective, we seek to design a photometric matcher and a geometric matcher capable of matching very wide-baseline images. The geometric matcher would be used to estimate the epipolar geometry without the knowledge of correspondence. The photometric matcher would help in initializing a set of putative correspondences

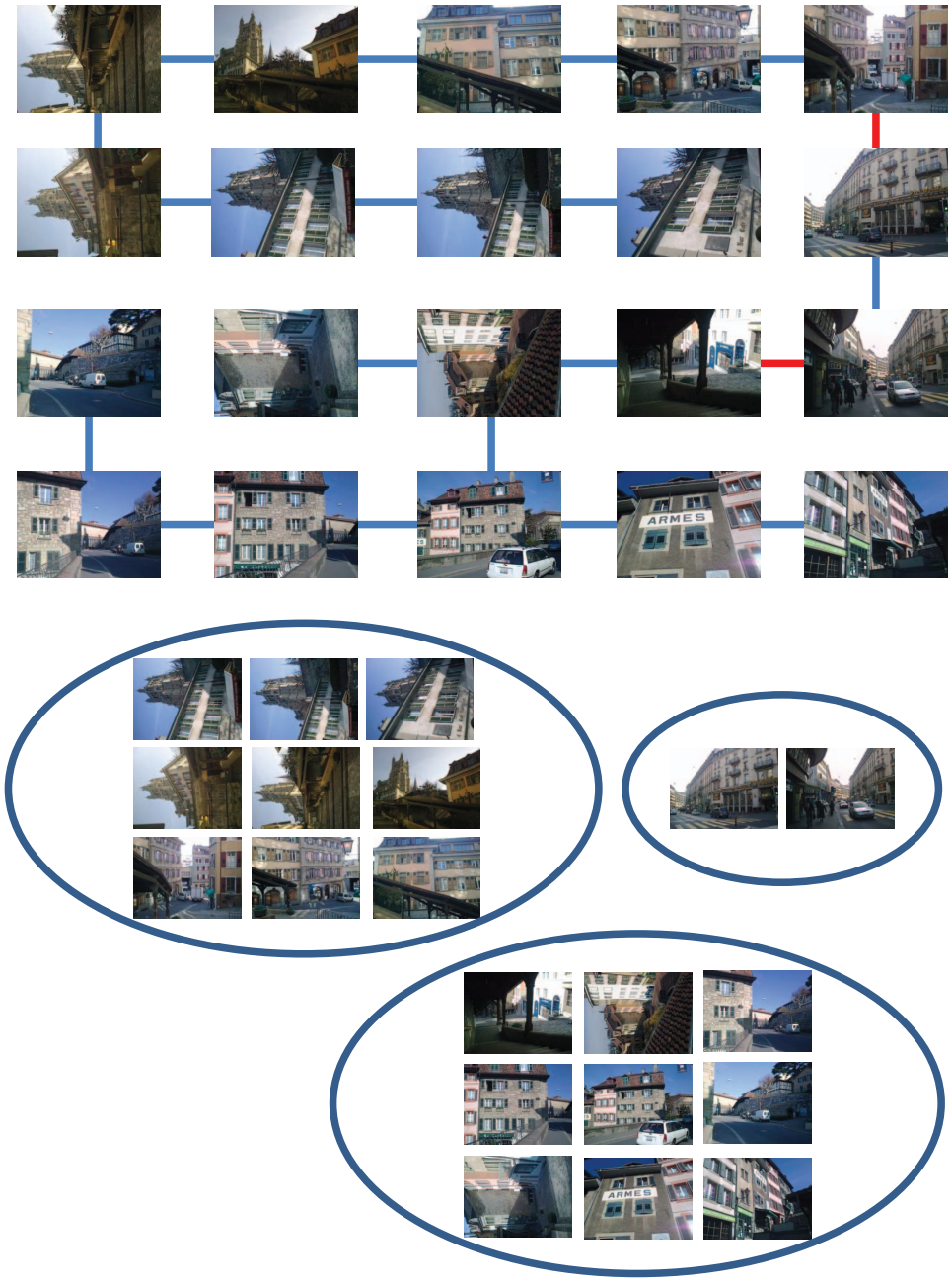


Figure 1.3: An organized collection of images. Note that organized collections are visually more satisfying.

and making the task simpler for the geometric matcher so that it only needs to identify the correct correspondences in the putative correspondences. The correct correspondences can then also produce the correct epipolar geometry. Geometric matcher has two issues that need to be addressed. First is that the geometric matcher is the bottleneck step in the pipeline. Thus, we want to reduce the time taken by the geometric matcher to efficiently match a pair of images. Second, factor is degeneracy. A sample from a data is degenerate if the sample does not represent the entire data. Our geometric matcher should be able to yield non-degenerate results.

2. Our second objective is to propose a geometric match score and a photometric match score that is a close approximation to the computationally expensive geometric match scores. The photometric match score would be produced using photometrically weighted putative correspondences that would be an approximation to the geometric match score that is produced using geometrically weighted putative correspondences. A task is to decide a threshold on the geometric and photometric match scores to capture actual visual connectivity between images.

3. Our third objective is to design an efficient algorithm for discovering connected components of visually connected images in wide-area sparse-view datasets with small number of expensive geometric estimations and still yield results similar to that with all pairs geometric scores. Since, the geometric match score estimation is computationally the bottle-neck step, we want to minimize the number of times it is used by using the photometric match score for all image pairs to guide the usage of the geometric matcher.

4. Our fourth objective is to automatically generate geometric walk-throughs in an image collection. The connected components are organized in linear (path) arrangement or planar (tree) arrangements as basal graph structures as shown in Figure 1.4. Path is a linear organization and tree is a non-linear organization

without a loop. Paths allow visualization of the collection as a video. A tree structure can be mapped to a 2D plane layout visualization of a collection.

5. Our fifth objective is to detect noisy location (GPS) and orientation (magnetometer) tags using pair-wise geometry estimates. GPS can be noisy for various reasons like ionosphere and troposphere delays, signal multi-path, receiver clock errors, orbital errors, satellite geometry/shading and intentional degradation of the satellite signal [4]. We need to identify reliable epipolar geometry estimates from matching images to use them to detect noisy tags. We seek to make this process computationally efficient by suggesting candidate image pairs for geometry estimations that are likely to match by photometric pre-filtering. Next, relative orientations and unit translations between pairs of cameras are found. Finally, we seek to use the reliable geometry estimates for estimation of rotation and unit translation between a pair of images to detect the error in sensor based tags.

1.2 Approach and Broadly Related Works

The problem of visual geometry based image organization including the wide-baseline version we are addressing involves four basic steps: first is feature extraction, second is feature matching and image matching, third is multi-view clustering and formation of minimal and foundational graphs in each cluster and the last stage is expansion of these graphs for various applications.

We are interested in the problem of organization of wide-area sparse-view datasets to leverage the geometric information in them. In this dissertation, we focus on the applications like geometric walk-throughs and detection of noisy position and orientation tags. Current research seeks to improve the accuracy of matching and to reduce the time and memory

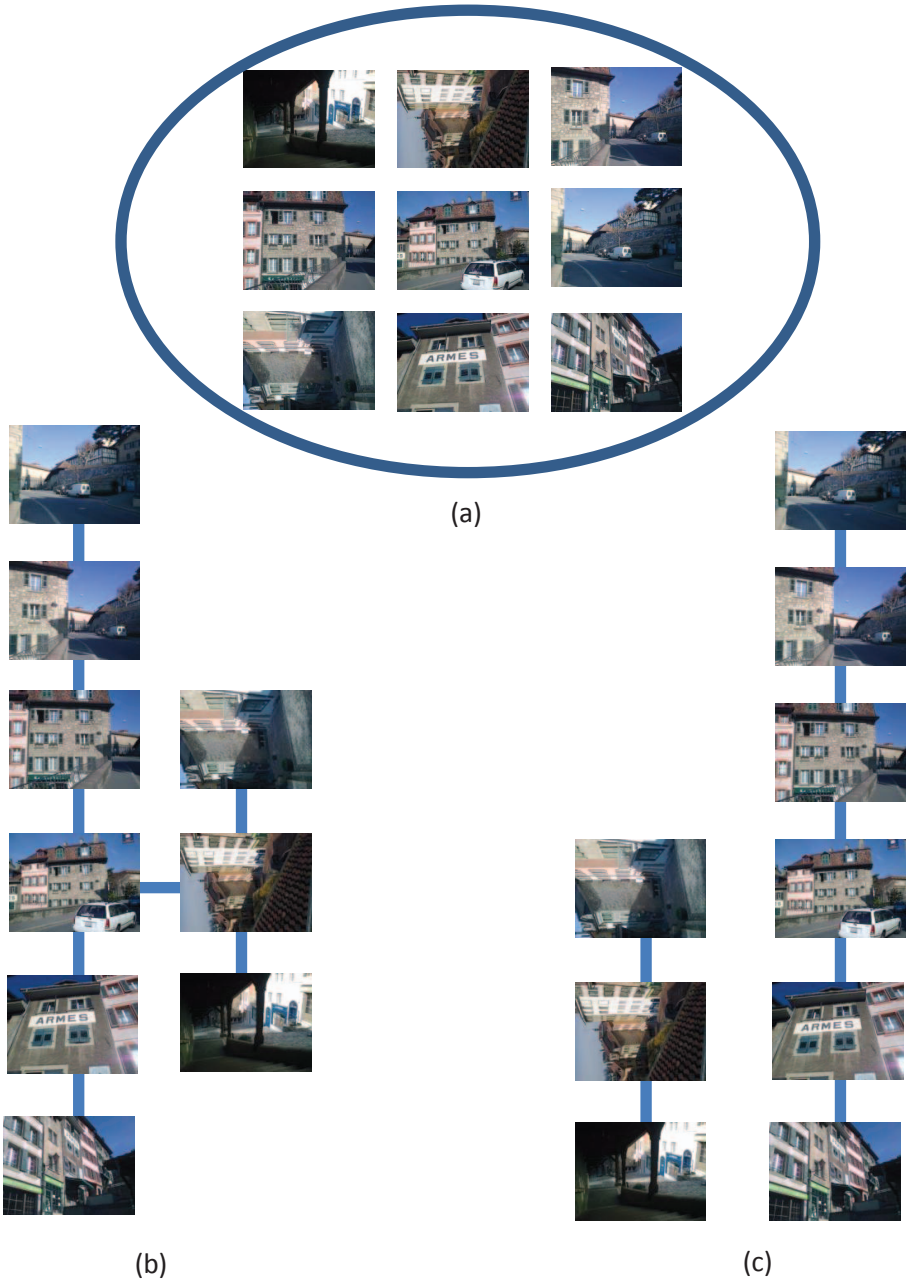


Figure 1.4: a) A set of connected images, b) a basal tree graph, c) basal path graphs. Note that a basal tree graph might not have a connected basal path graph corresponding to it.

usage of the entire pipeline of the four steps. Next, we discuss our approach and contrast it broadly with related works. More detailed discussions appear later in chapter 2.

First in the pipeline is feature extraction which is also one of the slowest steps in the geometry based image organization pipeline. This cost is linear over the number of images. Features that can be matched accurately are the best features. Among all kinds of features, point features have been most successfully used for matching specially for geometry based image organization. Among all point features, SIFT [77] features are most popular nowadays.

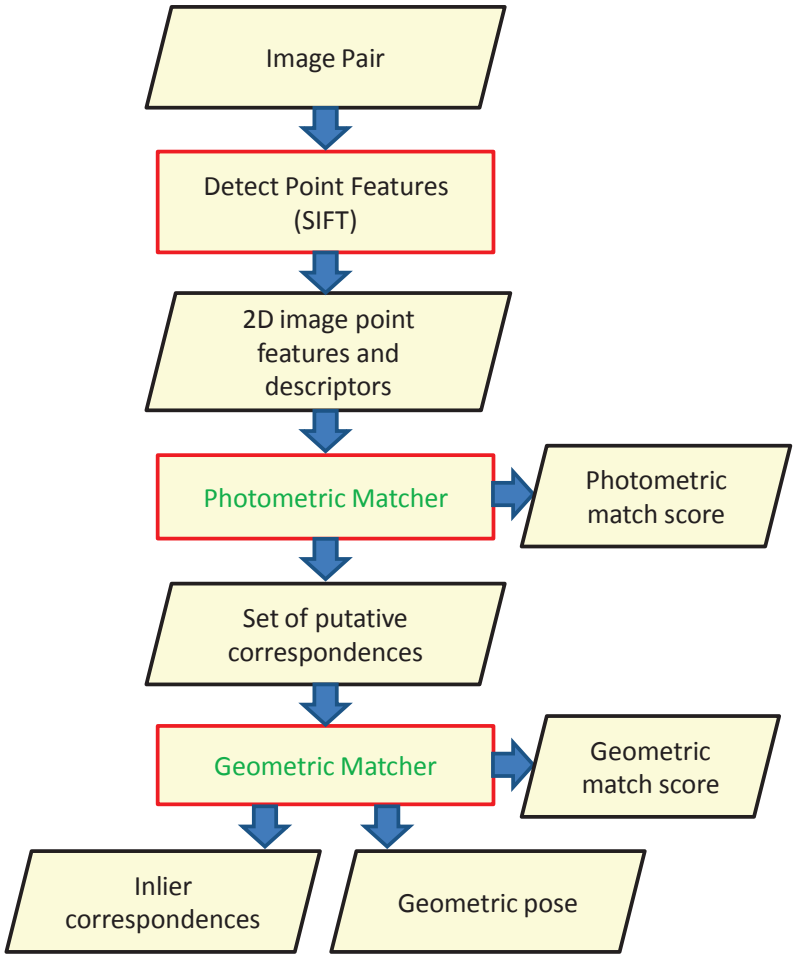


Figure 1.5: A flowchart showing how photometric and geometric match scores are produced by SIFT point feature matching between image pairs.

Second is the problem of feature matching, that is to find the corresponding points [79, 107] between a pair of images that can also be used to estimate the geometry of the associated cameras. Feature matching can be of three types: photometric, geometric or a hybrid of both photometric and geometric. Photometric constraints are weak constraints based on local appearance of a point feature. The strongest constraint on feature matching is the geometric constraint. The geometric constraint is that the corresponding points in two images must be projections of a unique 3D point, and thus, all corresponding points should constrain the geometry (position and orientation) of the two associated cameras. However, geometric matching is very expensive compared to photometric matching. If only the location of the 2D points in the images is used then the feature matching is purely geometric. If only the local descriptors of each feature are used, then the feature matching is purely photometric. If both the location and descriptors are used, then the matching is hybrid. Hybrid strategies of matching are most popular in the state-of-the-art. In hybrid methods, photometric matching is done first but all photometrically established correspondences do not necessarily satisfy the geometric constraint. Correspondences that do not satisfy the geometric constraints imposed by the highest majority of correspondences are discarded. This is best done in a random sampling framework where it is expected that after a number of trials, a sample of minimal points with all correct correspondences would be drawn producing an epipolar geometry supported by all the correct correspondences. RANSAC [47], MAPSAC [125], guided-MLESAC, M-estimator [33], pbM-estimator [103], PROSAC [37] and ORSA [86] are the few of the popular matching algorithms among others in the state-of-the-art. Image matching scores can be generated using the rate and number of inliers in putative correspondences identified in the feature matching process as shown in Figure 1.5. The parallelogram boxes in Figure 1.5 denote input or output. The rectangular boxes indicate algorithms. The red rectangular boxes with name of algorithm in green are our contributions.

We propose a 'hop-diffusion' Monte Carlo Markov Chain (MCMC) [27] strategy and a hybrid matching algorithm Balanced LOcal and Global Search (BLOGS) that is better

than the state-of-the-art in terms of speed and accuracy. First, we propose an epipolar geometry estimation algorithm for a pair of cameras separated by a wide-baseline. As described earlier, wide-baseline refers to large distance between the optical centers of the two cameras. Matching wide-baseline images is difficult because the same features might not get detected in both images and even if they do, their local appearance might vary a lot due to geometric transformations and changes in illumination between the images. Presence of repeated patterns makes the problem even more difficult. We calls our algorithm, BLOGS (Balanced LOcal and Global Search). Instead of random sampling done in most of the state-of-the-art algorithms, BLOGS does two kinds of guided sampling - local and global. Local guidance is based on a probability distribution defined by the best epipolar geometry estimated so far by the algorithm and global search is based on a probability distribution defined 'apriori' based on photometry. So we have two simultaneous processes: global hops and local diffusions. In the algorithm, local diffusions operate as a Markov Chain and if the global hops produce a better result, local diffusions start around that result. We call this strategy 'hop-diffusion'. BLOGS has a novel degeneracy measure based on the spectral scatter of a pair of correspondences. BLOGS performs best with an equal mix of hops and diffusions. BLOGS is 10 times faster than the competing state-of-the-art algorithms for same accuracy and takes the lowest time for a single iteration and the time per iteration is constant on increasing the number of iterations unlike other algorithms. BLOGS is also better in identifying more number of inliers within a threshold in a given number of iterations.

Third in the pipeline is the problem of multi-view clustering that is to identify connected components in a graph with image vertices and edges connecting images overlapping in scene content. The first step of this problem is to estimate this graph. Multi-view clustering can again be of three types: photometric, geometric or a hybrid of both photometric and geometric [13, 68, 75, 115]. Photometric methods compute photometric match scores between all pairs of images exhaustively. GIST features [45, 95] and Bag-of-Words (BoW) [15, 45, 92] of SIFT features are used in the state-of-the-art to produce

photometric scores. Geometric methods compute geometric match scores between all pairs of images exhaustively. However, the quadratic cost of estimation of geometric scores between all pairs of images is infeasible although it yields the most desirable results. The challenge is to minimize the number of geometric score estimations and still yield results as if geometric scores for all pairs of images were estimated. This challenge is answered by using hybrid methods for multi-view clustering where expensive geometric scores are opportunistically estimated using guidance done by cheap all-pair exhaustive photometric estimates. In the state-of-the-art, k photometric nearest neighbors of each nodes are retained in a graph for geometric score estimations. This has been the most commonly used or possibly the only algorithm used in vision for multi-view clustering to the best of our knowledge.

We collectively call minimal and application-wise foundational graph structures as basal graphs. In the state-of-the-art, basal graph structures are sought to establish connection between images in a collection. Snavely [115] proposed a minimal and foundational graph structure for 3D reconstruction called 'skeletal sets' (maximum leaf spanning trees). Thus, 'skeletal sets' proposed by Snavely is also a basal graph. In Snavely's algorithm, 3D reconstruction is first done over the internal nodes of the graph. Later, incremental reconstruction is done over this initial result using the leaf nodes. Unlike, narrow baseline image collections used in Snavely's work for 3D reconstruction, we are interested in leveraging the information in wide-area sparse-view datasets for other applications like geometric walk-throughs and detection of noisy position (GPS) and orientation (magnetometer) tags. We propose two other kinds of basal graph structures for geometry based organization of wide-area sparse-view datasets targeted towards these applications.

Our method is a unified approach involving clustering as well as seeking the basal tree graphs simultaneously using minimum number of geometric estimations. We use BLOGS to estimate geometric match scores and propose computationally cheap photometric match scores called Cumulative Correspondence Score (CCS) to approximate the computationally expensive geometric match scores. We propose an algorithm CODIMSEG

(COnnected component DIscovery by Minimally Specifying Expensive Graph) as a better alternative to k-NN approach used in the state-of-the-art for minimizing geometric estimations and minimally sacrificing accuracy compared to exhaustive geometric estimation between all image pairs. The CODIMSEG algorithm looks for a spanning tree in a graph of image vertices with most number of edges over a threshold on geometric match scores. CCS computation over all image pairs is very fast because we are able to use very small images and thus we can reduce the space and time. CCS gives a significantly more accurate approximation of the geometric match score than GIST as CODIMSEG algorithm identifies more edges above geometric match score threshold (that does not allow false positives) using CCS than GIST. Bag-of-Words is another method that yields results comparable to GIST [45]. However, performance of BoW is known to overly depend on training of the dictionary of words. Thus, we choose to compare CCS with GIST.

Fourth, in the pipeline is the expansion of the basal graphs (or skeletal graphs [115] proposed by Snavely) that connect images. Graph expansion is required because the basal graphs might be too sparse for many applications. In the state-of-the-art, graph expansion (better known as 'query expansion') is primarily done in two ways. One way [13] is by looking at transitive closures of the edges in the basal graph. Another way [68] is by looking at those edges first that maximally increase algebraic connectivity of the basal graph. However, both these strategies have some limitations. While looking for transitive closures is a good idea, looking for all transitive closures like done in the state-of-the-art is not. It is note-worthy that in a star-connected spanning tree, geometrically verifying all transitive closures would lead to exhaustive geometric estimations which is against the objective of minimizing geometric estimations. Thus, this would not be a good approach for any spanning tree that has nodes with high out-degrees. In the second approach, the possibility that images already connected via another image might be connected is ignored. The approach is more graph centric and the connectivity of the most connected images with the least connected images is sought and thus increasing the connectivity of the graph.

However, this method is likely to waste much geometric estimation as it completely ignores the distance of the two images in the basal graph.

We propose another algorithm for basal graph expansion. We also look at transitive closures of edges in the basal graph but only a minimal set of them that maximally connect the graph using a spanning forest again. In order to achieve this, we reuse our CODIMSEG algorithm to discover a spanning forest of all images directly connected to a vertex (image) under consideration in the image graph. All images in the basal graph are considered one at a time for graph expansion. After the graph expansion is done, it is used to produce basal path graphs for geometric walk-throughs. Basal path graphs are obtained by using approximate minimum Hamiltonian path through the expanded basal tree graphs. The expansion is required because a minimum Hamiltonian path is actually a degree-2-constrained minimum spanning tree and due to this constrain it is likely to split into parts. This happens because graph is a non-linear data structure and path is a linear data structure. Graph expansion reduces the number of splits. A geometric walk-through can also be done using just the basal tree graph traversal which is also a 2-approximation of a Hamiltonian path. However, the number of splits would be significantly higher. Expanded basal graphs are also used in one of our methods for detection of noisy position (GPS) and orientation (magnetometer) tags. This problem has never been solved in the state-of-the-art to the best of our knowledge. We propose a geometric voting scheme using image pairs connected in the expanded basal graphs to solve this problem. The vision based translation unit vectors and orientations between pairs of images connected in the expanded basal graph are used for the detection.

1.3 Algorithmic Contributions

Unlike state-of-the-art, we have targeted problems in more generally available image datasets which have not been collected with an application in mind and thus might or might not be visually connected. Even if they are visually connected, we do not assume that they would be connected by a narrow baseline. Thus, we address the problem of wide-baseline

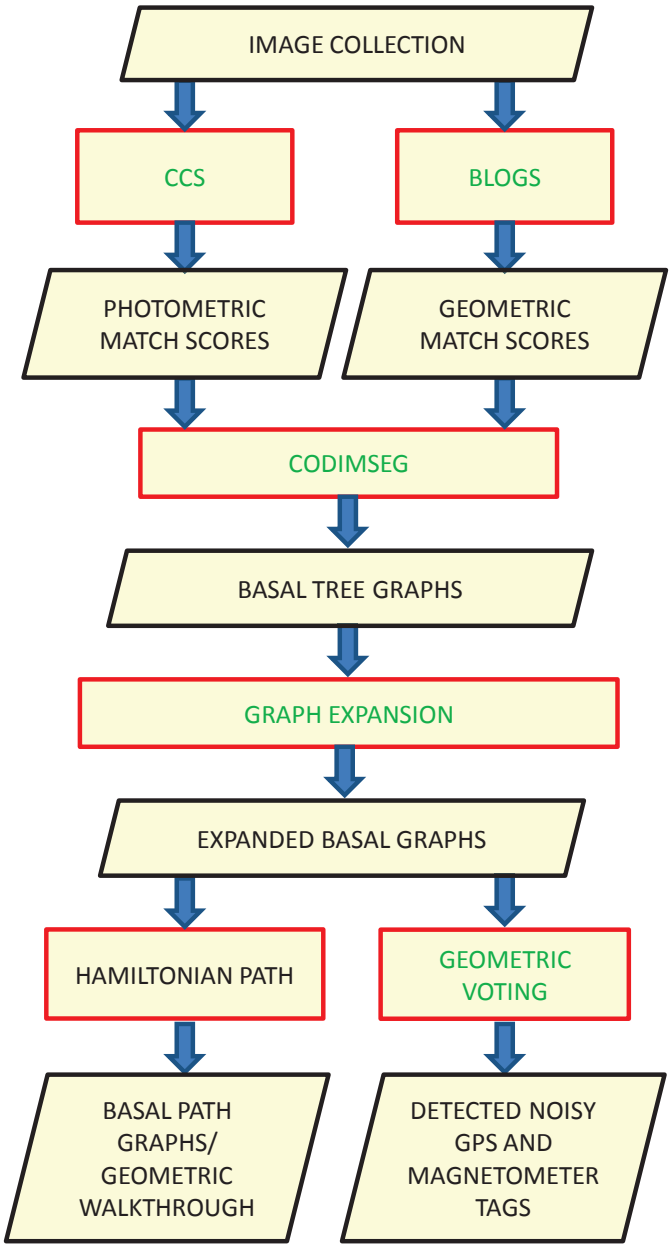


Figure 1.6: Overall flow, approach and contributions of the dissertation. The rectangular boxes with red borders are for the algorithms and the other parallelograms with black borders are for inputs and outputs. The algorithmic contributions are mentioned in green in the red boxes.

in image matching and the problem of organization of wide-area sparse-view datasets. We also looked at the problem of vision based detection of noisy image sensor based tags of position (GPS) and orientation (magnetometer) in wide-area sparse-view datasets.

We propose five novel algorithms in the dissertation as shown in the Table 1.1 and Figure 1.6. The red bordered rectangular boxes indicate algorithms. The other parallelogram boxes with black borders in Figure 1.6 denote input or output. The red rectangular boxes with name of algorithm in green are our contributions. These algorithms are :

1. CCS (Cumulative Correspondence Scores),
2. BLOGS (Balanced Local and Global Search),
3. CODIMSEG (COnnected component DIscovery by Minimally Specifying Expensive Graph),
4. basal graph expansion and
5. geometric voting.

These algorithms can be divided into three broad levels as shown in Table 1.1. The root level comprises of the problem of image matching for which we propose using CCS and BLOGS. The intermediate level comprises of the problem of generalized image organization using CODIMSEG and then expansion of the basal graph produced by CODIMSEG for variety of applications. Finally, the application level comprises of the problem of geometric walk-throughs in an image collection using approximate Hamiltonian path algorithms and detection of noisy position and orientation tags using geometric voting.

CCS is our photometric matcher and BLOGS is our geometric matcher which uses a novel Markov Chain strategy called 'Hop-Diffusion'. CODIMSEG is a novel general Markov Chain diffusion scheme for discovering connected components of a graph which is expensive to compute using a cheap approximation of the expensive graph. In our case, the geometric scores produced by BLOGS are accurate but expensive to compute and photometric scores

are approximate but cheap to compute. Basal graph expansion is a scheme for expanding the basal tree graphs or components connected by minimum spanning trees that are output from the CODIMSEG algorithm to make the graph more dense starting from basal. This dense graph is used for the problem of geometric walk-throughs using basal path graphs generated by approximate Hamiltonian path algorithms and for the problem of detection of noisy GPS and magnetometer tags.

The proposed algorithms significantly improve on the state-of-the-art. We show that our photometric score CCS is significantly better than GIST based score for image organization. Our photometric scores has a good correlation with geometric scores. Also, we show that our epipolar geometry estimation algorithm BLOGS is 10 times faster than the compared state-of-the-art algorithms namely MAPSAC, NAPSAC and BEEM. BLOGS performs well in degenerate conditions in which most of the point correspondences between the matching images come from the same region in the image. The time taken per iteration of BLOGS is the lowest among all the competing algorithms and it does not increase with the number of iterations unlike most other competing algorithms. BLOGS is also better in terms of identifying the correct correspondences within lower error thresholds. We compared CCS and GIST in our hybrid algorithm CODIMSEG which is a better alternative to k-NN as it does not need to specify k. k depends on the quality of the approximation which is generally not known. We compared the k-NN hybrid algorithm used in the state-of-the-art (with various values of k) with our hybrid algorithm CODIMSEG for minimizing the number of expensive geometric estimations in producing results similar to what would be produced if all-pair geometric matching were done. We found that our algorithm uses significantly less number of geometric estimates to reach the objective. In the noisy GPS and magnetometer tag detection algorithm, we found that the proposed algorithm performs well with respect to the ground truth. Problem of detection of noisy GPS and magnetometer tags has never been solved to the best of our knowledge. To test the accuracy of our GPS and magnetometer noise detection algorithm, we use very high-accuracy GPS coordinates from a differential sensor that is not commercially available in common multimedia devices.

Table 1.1: Table showing the contributions of the dissertation divided into three major levels mentioned with the challenges, contributions and the state of the art in each level.

Dissertation Problems	**Problem Category**	**What's new ? Challenge ?**	**Contributions**	**Representative State of the Art**
Matching image pairs	*Root level* (Building block: most used sub-problem)	Very wide-baseline Images	• Photometric matcher (CCS) • Geometric matcher (BLOGS)	• GIST, [Toralba 2002] Bag-of-Words (BoW) [Nister 2006] • MAPSAC, [Torr 2002] guided-MLESAC, [Tordoff 2005] pbM-estimator, [Chen 2003] BEEM, [Goshen 2008] ORSA, [Mosian 2004] PROSAC [Chum 2005]
Organizing image collections	*Intermediate level* (Generalized structuring of data for multiple applications)	Wide-area sparse-view datasets	• CODIMSEG : A Markov chain diffusion scheme for COnnected components DIscovery by Minimally Specifying an Expensive Graph. • Basal graph Expansion	Skeletal graphs [Snavely 2008], Image webs [Heath 2010], Iconic scene graphs [Li 2008], Photo-tourism [Snavely 2006], 'How do I organize my holiday snaps ?' [Scaffalitzsky 2002]
Detection of noisy image tags	*Application level* (A high impact consumer application)	Never solved (Identifying noisy image tags in wide-area sparse-view datasets)	•A geometric eigen- voting scheme (vision based voting for sensor based geometric measurements)	• Image localization [Khan 2006] • Bridging GPS outages in Mobile Mapping Vehicles (MMVs) [Roncella 2005]

1.4 Organization of the Dissertation

After the introduction, the dissertation is divided into five other chapters. Chapter 2 is on related prior works, chapter 3 on geometric matching, chapter 4 on image organization and basal graph structures, chapter 5 on noisy position tag and orientation tag detection followed by chapter 6 on discussions and conclusions. Each chapter has dedicated sections to discuss the background, the contributed algorithms, the experiments and the results. Figure 1.7 shows a flowchart and a chapter-wise distribution of the components in the flowchart. The flowchart has three main levels as shown in Table 1.1, the root level, the intermediate level and the application level, sandwiched between the input collection of images and the last chapter on discussions and conclusions. Root level comprises of chapter 3, intermediate level comprises of chapter 4 and application level comprises of parts of chapter 4 and whole of chapter 5. In Figure 1.7, the broad contents of each chapter and the chapter numbers are enclosed within blue-shaded blocks. The representative state-of-the-art algorithms are mentioned in elliptical regions in the blocks. In the flowchart, the red bordered blocks represent the algorithms and the black bordered blocks represent either the input or output. The novel algorithms are mentioned in green color in the red blocks.

In the first level, the flowchart shows that photometric scores and geometric scores are produced by the CCS and BLOGS algorithms respectively. Note that geometric scores are expensive and exhaustive estimation of geometric scores between all image pairs in the collection is not done and thus the block for geometric scores is dotted. In the second level, the photometric scores and the geometric scores are estimated as per the need of the CODIMSEG algorithm which results in tree structured connected components of graphs with image vertices which we call the basal tree graphs. The basal tree graphs are sparse and to use them for applications, they are expanded by verifying image connections nearby to the graph. In the third level, approximate Hamiltonian paths are found in the expanded basal tree graphs to produce basal path graphs. Geometric voting is done using the expanded basal tree graphs to detect noisy position and orientation tags. Finally, the dissertation

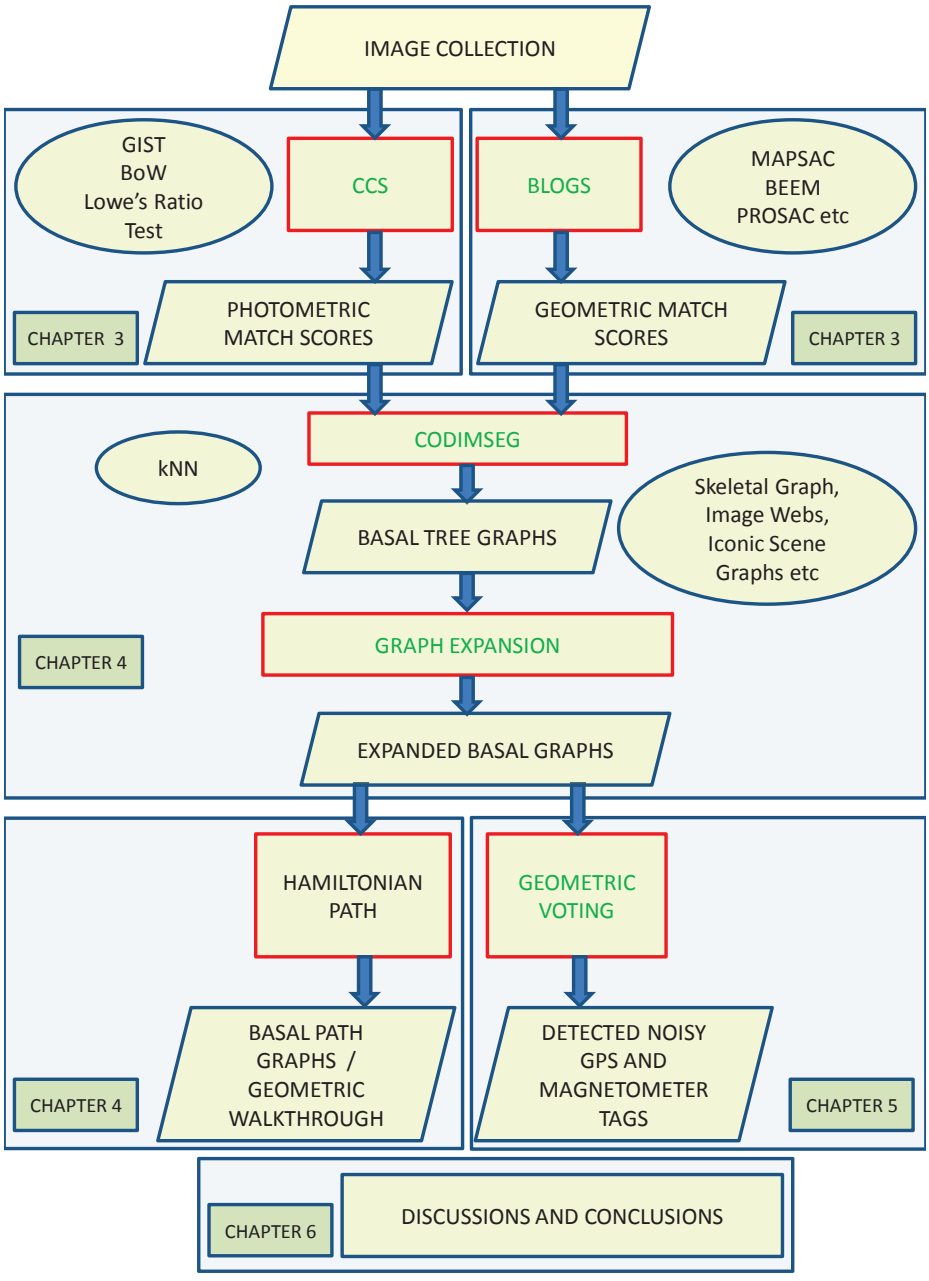


Figure 1.7: Flowchart showing the chapter wise flow of the dissertation. An image collections is an input.

reaches its conclusion after a discussion on the broader impact of the dissertation and possible future works. To assist the readers, a list of all the notations used in the dissertation can be found in the Appendix B among other Appendix chapters.

Chapter 2 Related Prior Works

In the state-of-the-art [53, 93, 94, 108], multiple narrow-baseline image datasets are used for applications like CBIR (Content Based Image Retrieval), 3D reconstruction and geometric walk-throughs. In contrast, we are interested in leveraging the geometric information in wide-area sparse-view datasets for applications like geometric walk-throughs and detection of noisy position and orientation tags as shown in Table 2.1. Unlike the state-of-the-art, we look at the problems in more general datasets for which no information is known a priori.

In the state-of-the-art, most algorithms have either shown results on very small datasets of less than hundred narrow-baseline images that were pre-collected to be of the same region, or they have used special processors like GPGPUs (General Purpose computing on Graphics Processor Units) [13, 50, 52, 68] or thousands of compute nodes in a much larger scale of up to millions of images. We are interested in the problem with additional constraints of memory as well as processing power of a simple personal computer or a hand-held device that has its own processor, so that large scale image organization can be realizable on future hand-held devices without connecting to superior machines. Table 2.1 shows the problems addressed in the dissertation in contrast to the state-of-the-art applications and approaches broadly.

2.1 Photometric Image Matching

Image matching based on appearance is called photometric image matching. Photometric image matching imposes weak match constraints and thus in general takes less time to match a pair of images. Photometric matching schemes that are commonly used in the

state-of-the-art are SIFT feature matching using Lowe's Ratio Test along with approximate nearest neighbor using KD-Tree [30, 114, 115], bag-of-words [15, 45, 92] and GIST [45, 95].

Lowe [77] proposed the SIFT features and a test to match the descriptors to faithfully produce good correspondences. He suggested that the ratio of the second-best match and the best match should be less that 0.6. This test is called the Lowe's ratio test. The Lowe's ratio test is very selective and often rejects many good correspondences as well and still does not guarantee absence of wrong correspondences. Due to very selective nature of the test, very few correspondences are able to pass it specially for wide-baseline image pairs. Thus, large images are needed to get sufficient number of correspondences to pass the test. Furthermore, on using large images, due to high number of descriptors per image, not only the feature extraction time increases, but also the feature matching time increases. The feature matching time is $O(n^2)$ where n is the number of features per image. Since, we need only the best match and the second best match, this complexity can be reduced. However, we need to use an approximate nearest neighbor algorithm. Some accuracy might be lost in the approximation of the best match and the second best match. KD-tree is an approximate nearest neighbor algorithm used for high dimensional data like SIFT that has 128 dimensions. KD-tree is formed using descriptors of one of the images to be matched and the descriptors of the other image are used to traverse the tree and match. Median along the direction of the highest variance is chosen to divide the tree in to two branches and so on for the sub-trees. Thus, the complexity of matching using KD-tree is $O(n \log n)$.

This method has several advantages and disadvantages. One of the advantages of using this method is that the percentage of correct correspondences found by matching SIFT descriptors and performing Lowe's ratio test is high. Thus, it takes less number of random sampling iterations to get a correct estimate of epipolar geometry. It is very commonly used in small scale Structure-From-Motion problems because not just the image similarity can be computed based on the number of correspondences using this method but also the correspondences are known.

Table 2.1: Table showing the problems addressed in the dissertation among other state-of-the-art problems and applications. The table broadly classifies various applications.

<table>
<tr><th rowspan="2">Applications</th><th rowspan="2">Primary Objective</th><th rowspan="2">Dominant Approach</th><th rowspan="2">Expected Nature of Ground Truth</th><th colspan="3">Dataset Type</th></tr>
<tr><th>Prevalent Scale</th><th>Narrow Baseline</th><th>Very Wide Baseline (Geographically Sparse)</th></tr>
<tr><td>Image Retrieval</td><td>Scalability</td><td>Fast global photometric matching
+
Weak geometric verification</td><td>Geometric</td><td rowspan="4">Hundreds of thousands of images using GPGPUs or clusters with thousands of nodes

OR

Hundreds of images on a single CPU</td><td rowspan="2">Lots of works</td><td>Global photometric matching and weak geometric measures do not identify very wide-baseline matches.

Please refer to future works in the dissertation.</td></tr>
<tr><td>3D Reconstruction</td><td rowspan="3">Accuracy</td><td rowspan="3">Local photometric matching
+
Strong geometric verification</td><td rowspan="3">Manual</td><td>Not visually appealing if done.</td></tr>
<tr><td>Navigation/ Geometric Walk-throughs</td><td>Few Works</td><td>Our Work
(Never done to the best of our knowledge)</td></tr>
<tr><td>Noisy Geometric Tags Detection</td><td colspan="2">Our Work
(Never done to the best of our knowledge)</td></tr>
</table>

However, this method does not scale very well for datasets of large size because it is expensive in memory and time due to use of large images that produces more descriptors needing more memory and more time to compute and match them. Due to memory and time constraints, new methods of matching like bag-of-visual-words and GIST got introduced into the image matching pipeline.

Bag-of-Words [14, 45, 70, 92] model in computer vision has drawn inspiration from text mining. In text mining, two texts can be matched using the frequency histogram of words appearing in the text. This motivated vision researchers to match images using similar method. The first problem is to create a dictionary of visual words. To build a dictionary of visual words, a training dataset of images is used and the SIFT features in all the training images are put together (in a bag), and a k-means clustering of the SIFT features is done to form visual words. A more advanced method is to do a hierarchical k-means clustering leading to formation of a vocabulary tree. After formation of the dictionary of visual words, each image in the test dataset is represented as a frequency histogram of these visual words. There are different ways of making the frequency histogram like binary (denoting presence or absence of words), Term Frequency (TF) and Term Frequency Inverse Document Frequency (TF-IDF). TF indicates how many times a visual word appears and DF indicates how many times a document with that word is found. IDF means weighing inversely using DF. TF-IDF is given by TF$\times$ log(-DF). TF-IDF is known to give the best results. Inverse Document Frequency is maintained by using inverted files. These frequency histograms are then matched with the query image. A faster technique of searching a histogram is to use the inverted files [110] which quickly narrows down the search to few documents containing visual words in the query image. Another method is to use min-hash [40] which uses multiple random binarized histograms of images in test dataset. Then, common techniques of histogram matching like chi-square matching and histogram intersection can be used to rank the narrowed down search result.

GIST features [95] mean summary of an image. GIST features of an image are found by applying a bank of steerable filters on multi-scale pyramid of an input image. The

energy at each scale and each orientation averaged over a region of image form an element of the GIST descriptor vector. Thus, if an image pyramid has 4 scales and the bank of steerable filters have 6 orientations and the image is divided in 4×4 regions, then the GIST descriptor vector would be of length $4 \times 6 \times 4 \times 4 = 384$. The dimensionality of this vector can be optionally reduced to 80 using PCA.

2.2 Geometric Image Matching

Robust epipolar geometry estimations using a tentative set of correspondences is best done in a random sample and consensus [47, 64] framework. A guided sampling approach [24] may be used to speed up the search for the best fitting epipolar geometry which has the maximum inliers to it amongst the correspondences. The epipolar geometry problem has also been looked at by using 3 correspondences [82, 85] at a time by finding maximally stable extremal regions [81]. Epipolar geometry estimations using 2 SIFT correspondences [57] or 2 MSER-LAF correspondences [96] have been proposed, but they rely on extrapolation of a single correspondence into multiple correspondences to represent a homography or an affine region. Estimation of epipolar geometry from two independent homographies or affine regions [66, 85] is common. Also, there are geometric coding [139] strategies in photometric feature matching and weak geometric consistency [70] based strategies that promise good results, but they are not expected to be as good as geometric matching. There are many such epipolar geometry estimation algorithms [24, 33, 47, 125]. Some require high overlap between views and others can work with widely varying views, with little overlap [24]. We use the latter since many view pairs in our problem could be widely separated.

There are many other significant works on fundamental matrix estimation and structure from motion [62, 66, 69]. Different papers have brought forward different key aspects of the problem like objective function form, degeneracy and sampling strategy. Salvi *et al.* [17] compares the performance of fundamental matrix estimation algorithms classified as linear, iterative and robust algorithms. Many other robust algorithms [66] such as LMedS,

M-estimator [33], pbM-estimator [103], MINPRAN [116] have also evolved over the years. PROSAC [37] is another such algorithm that randomly samples from progressively larger sets of correspondences ranked in order of higher to lower similarity scores between SIFT features. [41] comes up with a good method of detecting planar degeneracies and estimating the epipolar geometry in presence of a dominant plane. [51, 72, 98, 130] also address the problem of degeneracy. [29, 44, 72] are few algorithms other than BLOGS [24] that have laid the epipolar geometry and correspondence problem in a probabilistic framework. An eigen-vector approach for feature correspondence, proposed by Shapiro *et al.* [109] was a classic approach towards this problem. Another interesting problem is to estimate pose and structure without the knowledge of focal length [55, 117], or to find the lens distortion and view geometry together [48]. If focal length is known, pose can be estimated using [91].

Maximum Likelihood Estimate SAmple Consensus (MLESAC) [129] models the residual error distribution of correspondences, given a candidate fundamental matrix as a mixture of Gaussian inlier error distribution and uniform outlier error distribution. These conditional probabilities due to individual residual errors are assumed to be independent. The product of all the conditional probabilities leads to a measure of likelihood of the correspondence set given the candidate fundamental matrix. For each candidate fundamental matrix, the inlier rate that maximizes the log likelihood is found by expectation maximization. MLESAC looks for the fundamental matrix that maximizes the likelihood of the putative correspondence set.

MAPSAC [125] and Guided-MLESAC [120] are two popular variants of MLESAC. MAPSAC is the Bayesian version of MLESAC that improves upon it by maximizing the aposteriori probability instead of likelihood. Guided-MLESAC extends on MLESAC by using prior knowledge of validity of correspondences.

There are three aspects of the MLESAC-school of approach that form the background against which we advance the state-of-the-art. First, is related to the models used for inlier and outlier correspondences. While the inlier error distribution can be quite confidently modeled as Gaussian, assuming that outlier errors exhibit uniform distribution

is arguable. The nature of noise might be quite structured such as in the presence of repeated pattern. Second, MLESAC does not assume any prior knowledge of the validity of a correspondence and all correspondences are given the same weight of validity for a single candidate fundamental matrix. Third, inliers are assumed to be mutually independent, but mutual independence of the inliers might be a not be a correct assumption. Our algorithm seeks to improve on these probable shortcomings of MLESAC to return a non-degenerate estimate of epipolar geometry.

The N-Adjacent Points Sampling And Consensus (NAPSAC) algorithm is motivated by the observation that inliers generally occur in close proximacy of each other. In our algorithm, we also do a local search. However, NAPSAC would tend to get stuck in degenerate solutions. While inliers might be proximate, such inliers would not lead to a good estimate of epipolar geometry.

Balanced Exploration and Exploitation Model (BEEM) search algorithm has a local search as well as a global search. While the objective function of this algorithm is same as RANSAC, the search mechanism is different. BEEM introduces a way to calculate fundamental matrix using just 2 SIFT correspondences. However, this method is feature dependent. It uses the dominant angle of SIFT features to produce 3 more correspondences from 1. BEEM draws samples based on SIFT [77] match scores and tries to avoid degeneracy by sampling one point from outside the support set of the parameter estimate. It also incorporates the local optimization same as in LO-RANSAC.

Unfortunately, quadratic cost of geometry based pair-wise image matching is impractical and thus it is also a bottleneck step in large scale 3D geometry based applications. Thus, exhaustive pair-wise image matching is avoided.

2.3 Graph Structures for Image Organization

There are many state-of-the-art image graph connectivity algorithms. One of the early works done by Schaffalitzky [106] was '"How do I organize my holiday snaps?"'. He

showed results of graph based geometrical connectivity in images on small datasets with densely spaced images. Later on, this research was aggressively progressed by Snavely [12]. He proposed the idea of image connectivity and large scale structure from motion [114] in the form of photo-tourism. However, the images needed to have significant overlap, unlike our case. Snavely [115] also proposed skeletal sets for structure from motion problems, but for highly redundant images. In contrast, we are looking at image graphs with sparse connections. The focus of most of the current vision algorithms has been to use highly overlapping densely sampled scene photo collections for large scale reconstruction or camera pose estimation. For instance, Snavely [12] in his work on photo-tourism uses tracks of more than 20 keypoints across multiple images that are consistent with pairwise epipolar geometries between consecutive views in the track. These kind of algorithms typically exploit the high overlap in scene content between closely spaced views and can have problems when the images are widely spaced in 3D space, i.e. camera positions are widely separated. In such collections, it might be rare to have more than two views of the same scene.

Skeletal Graph is an image graph structure proposed by Snavely referring to a minimum number of connected images over which 3D reconstruction if done would approximate the 3D reconstruction done over the entire dataset. Other images are later added to this skeletal set. First, image pairs with more than 20 correspondences passing the Lowe's ratio test are initially connected in a graph structure. It is either assumed that the graph is connected or the skeletal set is found for the largest connected component of the graph. The geometric edge weights are then found for all the connected pair of images in the image graph. The geometric weight is given by the trace of the positional covariance matrix of the 3D reconstruction for a pair of images connected by an edge in the initial graph. Next, in the geometry weighted graph, an approximate maximum leaf spanning tree is found. The non-leaf nodes in the tree form the skeletal set. Another constraint on the skeletal set is that the best weight is not compromised by more than a "stretch factor" in order to maximize leaves.

Iconic Scene Graphs is an image graph structure proposed by Li et. al [75] for image browsing and reconstruction over connected images. In order to find iconic scene graphs, first the images are clustered into connected sets using k-means clustering on the GIST descriptors of the images. The number of clusters that the image sets must be divided into is not discussed in the paper. However, it is certainly a crucial step. Next, geometric verification is done for all image pairs in each cluster and the image that is geometrically most connected within a cluster is accepted as an iconic image. Images that are not registered to the iconic images are clustered again and the process is repeated to form more iconic images. This is a two step process and does not have a stopping criterion. Next, iconic images are connected by considering k-nearest neighbors of the GIST descriptor matching or bag-of-words matching of SIFT features using a spanning tree.

Gosele [56] proposed the use of community photo collections for multi-view stereo. From here on, the size of datasets severely went up triggering the advent of CBIR as a preprocessing step for fast initialization in structure from motion problems. Agarwal [13] performed large scale 3D reconstruction of the city of Rome from an image collection from community websites. The contribution of this paper was the use of parallel and distributed algorithms. In this research, images of Rome were downloaded from social networks websites and reconstructed using 500 nodes of computing power. This research focused on the problem of large scale 3D reconstruction in feasible time using parallel and distributed algorithms. Initial matches are established using bag-of-visual-words based matching and the top k_1 nearest neighbors of each image are geometrically verified forming connected components. Singleton images are discarded and the connected components are merged by geometrically verifying next k_2 nearest neighbors of all images. k_1 and k_2 are set to 10 in their experiments. These choices in the algorithm are made without supporting them in theory. Next step is the query expansion step, in which transitive closure is verified, that is, all direct connections between images indirectly connected to another image through an intermediate image is verified and connected if they pass the verification. This process is repeated 4 times. After query expansion, the skeletal sets are identified and reconstruction

over the skeletal set is done first and later rest of the images is added to the skeletal set and a global bundle adjustment is done over each connected component. Similar research has been done by Furakawa [52] but is not limited to one city. Frahm [50] proposed the reconstruction of the city of Rome by using GPUs instead of cloud computing. Thus, we see that in the state-of-the-art, CBIR techniques and cloud computing are used to speed up large scale 3D reconstruction.

Image webs [68] aims at improving graph connectivity bottlenecks for organized image browsing and 3D applications. We also share similar objectives with them. However, we concentrate more on image browsing. Image-webs use Hessian-affine, Harris-affine and Maximally Stable Color Regions detectors and SIFT descriptors. Tentative correspondences are found by using KD-tree for finding Approximate Nearest Neighbors in descriptor space, followed by using the Lowe's ratio test. RANSAC is then used to find the correct correspondences fitting the dominant affine transformation. The union of intersecting affine covariant elliptical regions of key-points associated with correct correspondences forms the affine co-segments that connect a pair of images. Image-webs aim at discovering such connectivity for a large dataset of images. The image-webs algorithm works in two phases. It is known that all-pair geometric image matching is infeasible. The objective of the first phase is to discover connected components similar to what would be found if all pair geometric image matching were performed. The objective of the second phase is to connect all edges within a component that would have been found if all pair geometric matching were done. In the first phase, similarity scores between images are found using bag-of-visual-word model and the top $k = 25$ matches are considered in the decreasing order of the similarity scores. The candidate matches within a connected component of the evolving graph are skipped. The graph is formed by considering candidate matches that can potentially merge connected components which are basically spanning trees. This phase stops when the frequency of merges falls below a certain threshold or when the allotted budget of co-segmentation operations is exceeded. Thus, at the end of this stage, we have a spanning forest of connected components of images connected by affine co-segments. In the next stage, edges within

each component are discovered. For each image in the graph, $k = 25$ photometric neighbor edges are considered. For each component, a graph Laplacian is found and its Fiedler vector is used to rank edges to consider for affine co-segmentation. The second eigen-value denotes the algebraic graph connectivity. The Fiedler vector is updated after each edge addition in the graph. This method basically attempts at connecting the images with high difference in their eigen-centrality. This would quickly raise the centrality of the image with lower centrality, thereby increasing connectivity. However, no stopping criterion has been proposed for this phase. It would not be meaningful to go over all the edges proposed by the nearest neighbors.

In the state-of-the-art, photometric filtering of unlikely matches is done using either GIST or Bag-of-Words (BOW) of SIFT features or both [45]. GIST features are weighted combination of the output magnitude of many multiscale-oriented filters where the weights are set using PCA. GIST can be used to broadly classify the images based on the content. In the BOW strategy [92], features descriptor vectors of training images are stored together and a hierarchical k-means clustering is done on all these descriptor vectors forming a dictionary of visual words (cluster centers) in the form of a vocabulary tree. Inverted file and min-hashing are considered good representative data structures for vocabulary tree for indexing images [15]. Query images are represented through vocabulary trees as a histogram of node visits. In the test database, candidate matches are ranked according to the similarity of the candidate image histogram in the database and the query image histogram. GIST has been shown to produce results comparable to BOW in [75]. However, GIST is not scale invariant and BOW requires a trained dictionary and undergoes quantization loss [97] in the descriptors due to k-means clustering. Thus, in this research, we propose a CCS(Cumulative Correspondence Score) measure for photometric filtering using point feature correspondences, use it with SIFT features and compare it with GIST's performance on the different datasets.

Other related works are on landmark recognition [32, 134, 138]. Landmark recognition also uses the geo-tags to do a geo-clustering and uses geo-tags to refine the noisy

landmark names in travel guides. However, we do not use geo-tags for clustering in our research rather we investigate the reliability of GPS information in a following research work.

2.4 Multimedia Position and Orientation Sensors and Vision

Detection of noisy GPS and magnetometer tags has not been addressed in vision research to the best of our knowledge. However, GPS and vision sensor fusion has been researched in the context of many applications like vision aided inertial navigation for flight control [133], localization of a query image in a dataset of GPS tagged images [73] [135], bridging of GPS outages in Mobile Mapping Vehicles [67, 101]. In all such applications, the visual connectivity among images is dense in terms of the 3D scene they capture. In our research, we face the challenge of sparse visual connectivity across views.

In both [73] and [135], algorithms for recovering the position of a camera to ascertain where a query image has been taken, given a dataset of GPS tagged images have been proposed. In [73], this is accomplished by finding the position of the epipole of the query image in a given route panorama. In [61], an algorithm to ascertain the position of video frames captured by a mobile camera is proposed, given images with GPS tags of some landmark locations along the route the camera follows. In [135] and [61], the position of the query image is ascertained by triangulating with two closest reference camera views. In [101] and [67], algorithms for bridging of GPS outages in Mobile Mapping Vehicles has been proposed. In [101], IMU (Inertial Measurement Units) for orientation estimation is not used. In [67], land-based and air-borne frameworks for the problem are integrated.

Snavely[12] proposed the use of GPS pose as initialization to bundle adjustment [131] on datasets containing images that are very well-connected in terms of the 3D scene they capture. In this work, we investigate and find that in datasets acquired using modern mobile phones having cameras equipped with GPS and magnetometers, the estimates are not good enough to be accepted as initializations to bundle adjustment. So, our work

considers the problem that appear prior to that considered by Snavely. Which of the GPS and magnetometer tags can we reliably use?

Chapter 3 Geometric Image Matching

Geometric image matching refers to matching two images using geometric constraints between a pair of images. In computer vision, epipolar geometry estimation is a basic problem that is still heavily researched [34, 58, 74, 100, 119, 137]. The problem is hard if correspondences are unknown or if a putative set of correspondences are known but the inliers and outliers in the set are unknown. Knowledge of correspondence helps in estimation of motion and structure and knowledge of structure and motion helps in establishing correspondences. Thus, a coupled update strategy with random initialization is usually the approach for solving this problem.

Current research challenges on this problem involve situations with wide baseline [20, 24], occlusions, high scaling and rotation, which result in significant amount of features in the scene with no correspondence. Figure 3.1 shows example of an image pair dealt in this work. In this work, we are looking at such hard problems with as high as 90% outlier rate. The minimal set of correspondences, e.g. 8 correspondences, needed for epipolar geometry estimation is referred in this work as a 'correspondel'.

3.1 Background

For the perspective camera model, there are two types of geometry or motion models that are commonly used : 'epipolar geometry' and 'homography'. While homography motion model is used for planar scenes, epipolar geometry motion model is used for non-planar or planar scenes. For planar scenes, epipolar geometry estimate is equal to a skew symmetric matrix multiplied by homography. However, if for a non-planar scene, the

estimate of epipolar geometry corresponds to only a single plane, it is considered degenerate. In planar scenes, generally point matching is relatively very easy. Again, while 'homography' is used for multi-view panorama stitching, it is not used for multi-view 3D applications. In our research, we use the epipolar geometry model.

Given a good estimate of fundamental matrix [78], non-linear method called bundle adjustment [131] is commonly used to optimize the 3D structure, fundamental matrices and the intrinsic calibration parameters all at the same time. Dellaert *et al.* [43] proposes direct optimization of re-projection error based objective function using an MCMC strategy, but it too requires a good initial estimate of structure and motion. [36, 80] also have similar objectives. A good initial estimate of fundamental matrix makes the process more accurate and fast. Moreover, epipolar constraint is the strongest constraint on the search for epipolar geometry, although weakness of the epipolar constraint lies in the fact that it does not discern among correspondences along the epipolar lines. Other common constraints are uniqueness, similarity and proximity. Photometry based similarity scores are often used to establish a putative set of correspondences to initially bootstrap the search of correspondences and epipolar geometry that leads to the most meaningful 3D structure using the epipolar constraint.

The problem of epipolar geometry estimation and correspondence establishment in presence of wide baseline, large scale changes, rotation, occlusion and repeated patterns leading to high outlier rate has been addressed in this work. We present an algorithm (BLOGS) based on a novel hybrid MCMC strategy, which we call 'hop-diffusion'. In this work, this strategy is used to search for the non-degenerate epipolar geometry with the highest probabilistic support of putative correspondences. The 'best so far' samples are either accepted or rejected in each iteration of our 'hop-diffusion' strategy. The quality of the samples is evaluated using a combination of a Welsch's M-estimator and a new degeneracy measure that rule out degenerate configurations. The hop steps are large movements spanning across the correspondence space guided by a photometry based unconditional proposal distribution, which makes them global moves. Diffusion steps are small

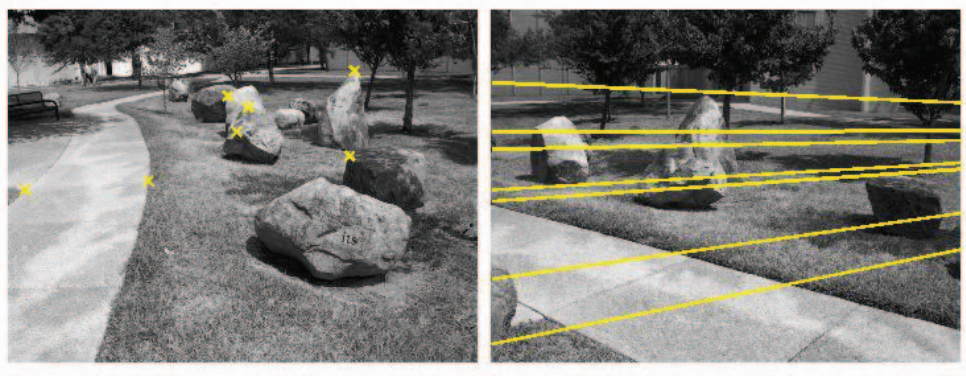


Figure 3.1: An example image pair used in our research with points on one image and epipolar lines on the corresponding image. Note the wide-baseline and the difficulty of the problem.

movements guided by a proposal distribution of likelihoods given by the Joint Feature Distribution(JFD) [128] conditioned on 'best so far' sample making it local to the 'best so far' sample. The algorithm's robustness with respect to its mixing parameter α (that sets proportion of hop moves and diffusion moves) and degeneracy parameter β has been studied empirically. The hop-diffusion framework allows handling upto 90% outliers even in cases where the number of inliers is very few. In practice, the contribution of our algorithm lies in higher precision and accuracy. BLOGS is compared with NAPSAC [88], MAPSAC [125] and BEEM [57] algorithm, which are the current state-of-the-art competing methods, on a dataset that has significantly more change in baseline, rotation, and scale than those used in the state-of-the-art. Not just is BLOGS able to tolerate very high outlier rates, but it also gives result of similar quality in 10 times lesser number of iterations as the most competitive among the compared algorithms.

We use SIFT features, random sampling and MCMC methods in this work and thus, next we give a brief background on these.

3.1.1 Scale Invariant Feature Transform (SIFT)

In the state-of-the-art, point features are detected either by Laplacian of Gaussian(LoG) [63], Difference of Gaussian(DoG) [77], or Determinant of Hessian(DoH) [21]. Laplacian of Gaussian is a second order derivative of Gaussian smoothed image, popularly known as Mexican Hat operator, used to get edges in an image. Difference of Gaussian and Determinant of Hessian are approximations of Laplacian of Gaussian. Harris corners, SIFT and SURF are few of the most popular point features detectors and descriptors. While SIFT uses DoG, SURF uses DoH. Images are matched using correlation based matching between detected corner image patches or by matching the descriptors of key-points. While many algorithms are popular for detection of features, SIFT is almost always the choice for description.

SIFT features use a scale-space pyramid to find out the scale invariant point features called key-points. As per Lowe's [77] paper, 4 octaves and 5 scales per octave are used to build the scale space pyramid. The octaves are re-sampled images and the scales are Gaussian smoothed images. The first octave is an up-sampled image with double the width and height of the original image and other two octaves are one-half and one-fourth the size. Up-sampling and down-sampling can be done using bilinear interpolation. Similarly, lower Gaussian scale in each octave is a fixed fraction times as much as the higher scale. For each octave, Difference of Gaussian(DoG) between images of consecutive scales are estimated. Difference of Gaussian is similar to Laplacian of Gaussian used for edge detection and is a much faster approximation to it. SIFT uses DoG for scale and Hessian for space to detect keypoints. Next, for each octave, after the 4 DoG images are obtained, key-points are detected at a particular DoG using 1 higher DoG image and 1 lower DoG image, that is, 3 consecutive out of the 4 DoG images. Key-points are extrema (minima or maxima) in the $3 \times 3 \times 3$ neighborhood surrounding the pixels in the middle DoG image. The associated scale of the key-point is stored. The key-points obtained are not exactly located at pixels, that is, it has sub-pixel locations. To find the exact key-point locations, a Taylor expansion

is done using pixel location around the approximate key-points. On differentiating and equating the Taylor equation to zero, the exact sub-pixel locations of the key-points are obtained. Next, if the magnitude of the key-point intensity found at the exact location of the key-point(put in the Taylor expansion) is below a threshold of 0.03 in the associated edge image, then the key-point is discarded. Also, if the key-point lies on a flat region or edge, determined by the ratio of the eigen-values (10 in Lowe's inplementation) of the Hessian matrix, it is discarded. The ratio of eigen-values can be quickly found without estimating the eigen-values by using the trace and determinant of the Hessian matrix.

Further, each key-point is described by a 16×16 window centered around the key-point and is divided into 16 4×4 windows. For each 4×4 window, the gradient orientation of each pixel is quantized into 8 histogram bins of 45 degrees and is weighted with Gaussian smoothed (pixels far from the key-point are weighed down) gradient magnitude. Thus, a key-point is described by a descriptor of length $4 \times 4 \times 8 = 128$. The features descriptors are normalized to a unit vector. Further, every key-point descriptor is assigned an orientation and a scale. The orientation is found by quantizing the gradient orientation of all the pixels of the 16×16 window centered around the key-point into 36 bins of 10 degrees and weighting them by absolute gradient magnitudes. The dominant bin in the histogram denotes the orientation. If there are other bins higher than 80% of the dominant bin in the histogram, then the keypoint is duplicated with other such orientations as well. The orientation of the descriptor is relative to the orientation of the keypoint to make it rotation invariant. Illumination invariance is introduced by thresholding values greater than 0.2 in the descriptor and making them equal to 0.2 and normalizing it again to a unit vector.

Some variants of SIFT are PCA-SIFT [71] and GLOH [84]. In both PCA-SIFT and GLOH, the descriptor is different from SIFT. In PCA-SIFT, patches of 41×41 pixels centered around the key-point are considered giving 39×39 horizontal and vertical gradients relative to the dominant orientation of the key-point. This gives a $2 \times 39 \times 39$ vector. This vector is normalized to get a unit vector. Using PCA, dimensionality of the vector is

reduced to 36 and re-normalized. PCA-SIFT is sometimes less discriminatory than SIFT and is computationally more expensive.

In GLOH, the pixel locations around a key-point are quantized into log polar histogram bins instead of having square cells with Gaussian weight like SIFT. There are 3 bins in radial direction of radius (6, 11 and 15) and 8 bins in angular direction. The first radial bin is undivided and thus total bins are 17. In each bin, gradient orientation is quantized in 16 bins. Thus, total size of the descriptor is $16 \times 17 = 272$. The descriptor is normalized to unit-vector and the size of the descriptors is reduced to 128 using PCA and renormalized. GLOH is computationally more expensive.

Both PCA-SIFT and GLOH require a PCA training step. The covariance matrix for the PCA is trained over collected images patches.

3.1.2 Eight-Point Algorithm

The eight-point algorithm [76] uses a Direct Linear Transform algorithm for estimation of transformation between at least 8 given correspondences between 2D points. The 2D points can be expressed as homogeneous coordinates and thus the transformation matrix should be a 3×3 matrix with at least 8 unknowns up to a scale. Thus, eight points are required to estimate the transformation matrix. The 9×1 eigen-vector corresponding to the lowest eigen-value produced as a result of eigen-decomposition of a matrix formed by a stack of eight correspondence tensors is reshaped as a 3×3 transformation matrix. However, this transformation matrix is not a fundamental matrix yet. Fundamental matrix has 7 degrees of freedom, 3 for rotation, 2 for translations up to a scale and 2 more for unknown focal lengths of the two cameras. To enforce this constraint on the 3×3 transformation matrix, a rank 2 constraint is applied to it by reconstructing the transformation matrix using only top two of its eigen-values giving the fundamental matrix.

Normalized eight-point algorithm [64] is a version of the eight-point algorithm that normalizes all the points before the Direct Linear Transform algorithm is used. The

normalization is done by translation of the point such that the centroid of all the points lie on the origin and scaling is done such that the mean distance of all the points from the centroid is $\sqrt{2}$.

Another version of the eight-point algorithm is the seven-point algorithm that uses theoretically minimum number of points to estimate the fundamental matrix that has a degree of freedom equal to seven. Instead of forcing a rank-2 constraint like the eight-point algorithm, the seven-point algorithm uses two eigen-vectors corresponding to the two lowest eigen-values and uses a Lagrange's multiplier between the two transformation matrix thus obtained. A rank-constraint is applied to this combination leading to a cubic equation in terms of the multiplier. Solving the equation would give either 1 or 3 real values of the multiplier and thus 1 or 3 distinct fundamental matrices are produced. The seven point algorithm can also have a normalized version.

For more details, please refer to [66].

3.1.3 Random Sample and Consensus Class of Algorithms

RANdom SAmple Consensus (RANSAC) [47] is a robust algorithm [83] used to identify inliers and inlier model in a noisy data. It has many variants. Some of which are [28, 35, 38, 39, 49, 89, 90, 99, 122, 123, 126]. The algorithm works by drawing multiple random samples of minimal model size and verifying all the data with respect to the models thus formed. The sample that has highest consensus in the data gives the model chosen and the inliers to this model are accepted as inliers in the data.

However, there are some critical issues that need to be taken care of when using RANSAC. The consensus is decided based on a threshold. If the threshold is too high or too low, the model chosen might not be best. Another important thing is that the number of samples that need to be drawn to get a sample free from outliers depends on the outlier rate in the data, but more samples might be needed if the data is degenerate in order to

come of the degeneracy. Degeneracy means that each data in the sample does not add information in the model because they carry the same information.

3.1.4 Monte Carlo Markov Chains

Monte Carlo Markov Chains (MCMC) [18, 54] is a popular learning strategy. Monte Carlo methods and Markov Chains are two different classes of algorithms and MCMC are a combination of both. Monte Carlo methods are random walk algorithms that do not attempt at making a directional choice but rather wait till the distribution of samples generated from the random process gets stationary. A Monte Carlo method starts by generating random samples and then analyzing the ones that obey certain properties to give numerical solutions to analytically difficult problems.

A Markov Chain is a sequence of random variables that show Markov property. Markov property is shown if the future state is dependent only on the current state and is independent of all past states. Three necessary properties of a Markov Chain are :

1. Ergodicity : All states are reachable from all other states in finite time.
2. Aperiodicity : State transitions should not be period.
3. Detailed Balance : Every transition is reversible, ensuring that the random process has reached a stationary distribution.

Sometimes the third property of Markov Chain is sacrificed or violated to gain speedy results. Ensuring detailed balance ensures convergence and optimality, but takes a lot of time. Greedy methods generally accept sub-optimal results, but much earlier. There are greedy MCMC methods like Iterated Conditional Modes and hybrid MCMC methods like jump-diffusion. Our method is a greedy hybrid MCMC, which we call hop-diffusion. Our algorithm adopts a hybrid global hop and local diffusion mechanism in a MCMC framework.

Metropolis-Hastings [18, 54] is another popular MCMC method. The goal of this method is to simulate drawing random samples from a target distribution from which sampling directly is hard. In Metropolis-Hastings, the ratio of objective functions is multiplied with the ratio of effective proposal probabilities ensuring a reversible Markovian cycle by Bayesian probability inversion to attain a stationary distribution after a number of "burn-in" samples are generated. It is generally difficult to scale multiplying ratios in a way that one ratio does not dominate other. For example, a ratio of exponential terms is likely to dominate a ratio of linear polynomial terms in the overall multiplication result. In Metropolis-Hastings, the multiplication of ratios of objective function and ratios of effective proposal probabilities face this trouble as both the functions might be of different orders or categories, leading to undesirable result. Unlike, the Metropolis algorithm that draws samples from the objective function using the proposal distribution, we are just interested in keeping the 'best so far' sample.

The Iterated Conditional Modes [23] algorithm would in general start with an initial sample. All samples differing in at most one element from the current sample is found and the objective functions for all of them are evaluated. The next accepted sample is the one that locally optimizes the objective function. This algorithm overly depends on the initialization and is likely to get stuck in local maxima. The algorithm is similar to ours in looking for a nearby sample that is better than the current sample in more deterministic fashion than ours. In our algorithm, we are looking at local neighbor samples as well as global samples so that we are less likely to get stuck at local maxima.

Hop-Diffusion is a mix of random global hop moves and local diffusion moves that seek to optimize an objective function. Hop moves explore the space of parameters unbiased on any previous result and thus it a global exploration move. Diffusion moves explore the space of parameters based on the best parameter found so far by the algorithm and thus, diffusion is a local exploration move and exploitation move as well. In other words, the global hop search helps to arrive at different parameters and local diffusion searches are done to fine tune these parameters to see if a nearby solution is better. Coordination

between the hop moves and diffusion moves can be established using a parameter supplied to the algorithm or can be dynamically set as the algorithm proceeds. While local diffusion moves speed up the search process, they are more likely to get stuck in a local maxima. Hop moves being unbiased and global, can potentially help to get out of local maxima. Hop moves also upstages the platform by producing a better estimate that diffusion moves can compete and benefit from. We have investigated the behavior of our algorithm with different mix of hops and diffusions.

In the literature [118], many hybrid samplers have been reported. Jump-diffusion is one such hybrid sampler. Jump diffusion processes was introduced by Grenander and Miller [60]. Green's [59] paper more specifically deals with image processing and vision related research. Han *et. al* [46] has used jump-diffusion framework for range image segmentation more recently. While, hop-diffusion and jump diffusion are similar in principle, the major difference between the two lies in the fact that while jumps are between different sub-spaces, hops are done within one sampling space.

3.2 Our Approach: Balanced Local and Global Search

In this work, we come up with a new hybrid MCMC strategy, which we call hop-diffusion. We also come up with a novel method to detect degeneracy. These strategies can be applied in general. However, in this work, we have applied these strategies for epipolar geometry estimation. All the known aspects of epipolar geometry estimation have been addressed in this work and a thoughtful choice of optimizing criterion and search mechanism has been made. The framework of MCMC that we have described in the work performs better when searching for an estimate of epipolar geometry, which is shown by finding that our algorithm outperforms state-of-the-art algorithms.

The way we benchmark the performance of the state-of-the-art algorithms with our algorithm is also more meaningful and stringent than the state-of-the-art. We have hand marked ground truth on many test images on which we evaluate the epipolar geometry of all

competing algorithms. These hand marked ground truth points are not part of the putative set of correspondences. Thus, we have a clear separation of test and train, in some sense. Each algorithm is run 100 times on each image pair and the mean and standard deviation of root mean performance is reported. The algorithm that returns the least error in lesser number of iterations is of course the desirable one.

In our algorithm BLOGS, global searches are done using a distribution of based on photometry measures and local searches are done using Triggs' [128] Joint Feature Distribution that essentially imposes the epipolar constraint in a probabilistic way. JFD guided samples are also referred to as geometry guided samples. We randomly choose, at each iteration whether to draw a photometry guided sample or a geometry guided sample. Geometry guided samples are drawn from a distribution of conditional probabilities of putative correspondences conditioned on best known 'correspondel' so far. Thus, our guidance strategy necessarily involves a Markov Chain.

3.2.1 Problem Model, Notations and Mathematical Objective

Without the loss of generality, given any two images $\mathcal{I}_i$ and $\mathcal{I}_j$ from the collection $\mathcal{V}$, let the image with smaller number of detected features be $\mathcal{I}_i$ and let it contain n_i features such that $\mathbf{f_i} = [\mathbf{f}_i^1, \mathbf{f}_i^2, \cdots, \mathbf{f}_i^{n_i}]$. Let us add a NULL feature $\mathbf{f}^0$ to the feature set of the other image $\mathcal{I}_j$ with n_j number of features such that $\mathbf{f}_j = [\mathbf{f}^0, \mathbf{f}_j^1, \mathbf{f}_j^2, \cdots, \mathbf{f}_j^{n_j}]$. Any number of features in $\mathbf{f}_i$ can correspond to the NULL feature in $\mathbf{f}_j$. All other features in $\mathbf{f}_i$ should have one to one correspondence with features in $\mathbf{f}_j$. Mapping these n_i features to $n_j + 1$ features is the correspondence problem. This combinatorial problem is clearly NP-hard.

The above problem is reduced to a simpler problem by coming up with a putative set of correspondences based on photometry of the features. Then the problem is reduced to identifying the correct correspondences among these putative correspondences. The correct set of correspondence would give the best epipolar geometry and the inliers to the best epipolar geometry should only be the correct correspondences. All wrong correspondences

in the putative set of correspondences ideally should be outliers to the best estimate of epipolar geometry. However, the epipolar constraint constrains the correspondences along an epipolar line and thus cannot discern between two possible matches along the line.

Let $\mathbf{X}_{ij} = [\mathbf{x}_{ij}^1, \mathbf{x}_{ij}^2, ..., \mathbf{x}_{ij}^{|\mathbf{X}_{ij}|}]$ denote the putative correspondence set between images $\mathcal{I}_i$ and $\mathcal{I}_j$ where $\mathbf{x}_{ij}^k = \mathbf{v}_{ij}^k \otimes \mathbf{u}_{ij}^k$ is a 9 component tensor of match pairs $\{(\mathbf{u}_{ij}^k, \mathbf{v}_{ij}^k) | k = 1, \cdots, |\mathbf{X}_{ij}|\}$ given by $\mathbf{x}_{ij}^k = [x'x, x'y, x', y'x, y'y, y', x, y, 1]^T$ where $\mathbf{u}_{ij}^k = [x, y, 1]^T$ is the homogeneous coordinate of a feature point in image $\mathcal{I}_i$ and $\mathbf{v}_{ij}^k = [x', y', 1]^T$ is the homogeneous coordinate of the putative corresponding feature point in image $\mathcal{I}_j$. Let the number of putative correspondences between images $\mathcal{I}_i$ and $\mathcal{I}_j$ be $|\mathbf{X}_{ij}|$.

A minimal sample of s correspondences $\theta = [\mathbf{x}_{ij}^{d_1}, \cdots, \mathbf{x}_{ij}^{d_s}]$ is drawn from $\mathbf{X}_{ij}$ where $\mathbf{d}$ is a vector of indices of the s correspondences in the samples. Using this sample, we estimate the fundamental matrix $\mathbf{F}_{ij}(\theta)$. A fundamental matrix $\mathbf{F}_{ij}(\theta)$ must fit the constraint given by ${\mathbf{v}_{ij}^k}^T \mathbf{F}_{ij}(\theta) \mathbf{u}_{ij}^k = 0$ for all correspondences in θ. This constraint can be expressed as $\theta^T f = 0$ where $\mathbf{f}$ is a 9×1 representation of 3×3 matrix $\mathbf{F}_{ij}(\theta)$.

The problem is to find the fundamental matrix without the knowledge of correct correspondences in the putative set. Thus, we want to maximize the probability of the fundamental matrix given the putative set. In our work, we are looking to optimize the objective function $p(\mathbf{F}_{ij}(\theta)|\mathbf{X}_{ij})$ where its definition is given by

$$p(\mathbf{F}_{ij}(\theta)|\mathbf{X}_{ij}) = \frac{\mu(\mathbf{F}_{ij}(\theta)|\mathbf{X}_{ij})\omega(\theta)}{\int_\theta (\mu(\mathbf{F}_{ij}(\theta)|\mathbf{X}_{ij}))\omega(\theta)}. \tag{3.1}$$

$\mu(\mathbf{F}_{ij}(\theta)|\mathbf{X}_{ij})$ is the quality of the sample of s-tuple θ over the putative set $\mathbf{X}_{ij}$, and $\omega(\theta)$ is the 0/1 binary degeneracy measure of the sample θ. The quality of the sample is an estimate of the fitting of the fundamental matrix found using the sample θ on the putative correspondences. Effectively the objective function is designed so that the best non-degenerate sample is picked. In the objective function, the non-degeneracy measure weeds out the degenerate samples so that the fitting quality measure returns the quality measure of the non-degenerate samples.

After a number of "burn-in" samples of s-tuple θ are drawn, the one that maximizes $\mu(\mathbf{F}_{ij}(\theta)|\mathbf{X}_{ij})$ and has a $\omega(\theta)$ value of 1 is chosen as optimal θ^*. This fitting quality $\mu(\mathbf{F}_{ij}(\theta^*)|\mathbf{X}_{ij})$ or μ_{ij} in short, is basically a soft estimate of the number of inliers in the putative correspondence set. The geometric estimate of the rate of inliers in the putative correspondence set is given by

$$\phi_{ij}^1 = \frac{\mu(\mathbf{F}_{ij}(\theta^*)|\mathbf{X}_{ij})}{|\mathbf{X}_{ij}|}. \tag{3.2}$$

The geometric estimate of the number of inliers in the putative correspondence set given by

$$\phi_{ij}^2 = \mu(\mathbf{F}_{ij}(\theta^*)|\mathbf{X}_{ij}). \tag{3.3}$$

ϕ_{ij}^1 and ϕ_{ij}^2 collectively referred to as ϕ_{ij} is our geometric measure of similarity between two images. Whether images match or not can be judged using a threshold on both components of ϕ_{ij}. However, the number of iterations needed to get a value of ϕ_{ij} is high and this kind of geometric matching is expensive. Moreover, the $O(N^2)$ cost of estimation of ϕ_{ij} for all image pairs is infeasible for large values of N.

3.2.2 Hop using Photometric Proposal Distribution

The photometric proposal function is dependent on two primary attributes of each correspondence in the putative set. The first is what the similarity value is and the second is how distinct this similarity value is from that of the closest competing correspondence in $\mathbf{S}$. Many similarity measure have been discussed in [104]. In our case, the similarity is the reciprocal of distance between SIFT features in each image. Unlike Lowe's SIFT paper [77], we do not put a hard threshold on this distinctness and also use the similarity value and not just the distinctness. With each putative match pair, we associate a similarity measure, which we refer to as photometric weights in our work. Let the highest similarity in a row

and column of a match be ρ_{ij}^k. Let $\rho_{ij}^{k_r}$ be the second highest similarity in its row and $\rho_{ij}^{k_c}$ be the second highest in its column. We construct a weight w_{ij}^k for the correspondence x_{ij}^k given the photometry ρ as

$$w_{ij}^k = (1 - \exp^{-\rho_{ij}^k})^2 (1 - \frac{\rho_{ij}^{k_r}}{\rho_{ij}^k})(1 - \frac{\rho_{ij}^{k_c}}{\rho_{ij}^k}) \tag{3.4}$$

The measure we get is always between 0 and 1. All the multiplying terms individually range from 0 to 1. The first term is high if the similarity is high. The second and third terms are high if the closest competing correspondence in the row $\rho_{ij}^{k_r}$ and column $\rho_{ij}^{k_c}$ of $\mathbf{S}$ respectively are low compared to ρ_{ij}^k. The distribution of photometric weights in w_{ij}^k is used to draw a photometric correspondel sample.

3.2.3 Diffusion using Geometric Joint Feature Distribution

Triggs [128] proposed the concept of Joint Feature Distributions (JFDs) to provide a flexible and robust alternative to the strict and deterministic geometric constraints used for projective matching. In our context, we are particularly interested in two-camera 2D to 2D epipolar constraint. Simply put, JFDs are the joint probability distributions over the parameters of the corresponding 2D to 2D features. They summarize the statistics of a given set of correspondences and do not rigidly constrain them to a deterministic geometry. That is why they are an ideal formalism to account for small non-rigid distortions and errors that will inevitably be present in any camera. We use the Joint Feature Distributions to sample correspondel and guide the MCMC locally. The use of the conditional JFD alleviates the need for assuming that correspondences are independent of each other, a common assumption in many random sampling approaches for epipolar geometry estimation. We next summarize the development of JFDs. For more in-depth understanding of the concept, the reader should read the original article [128].

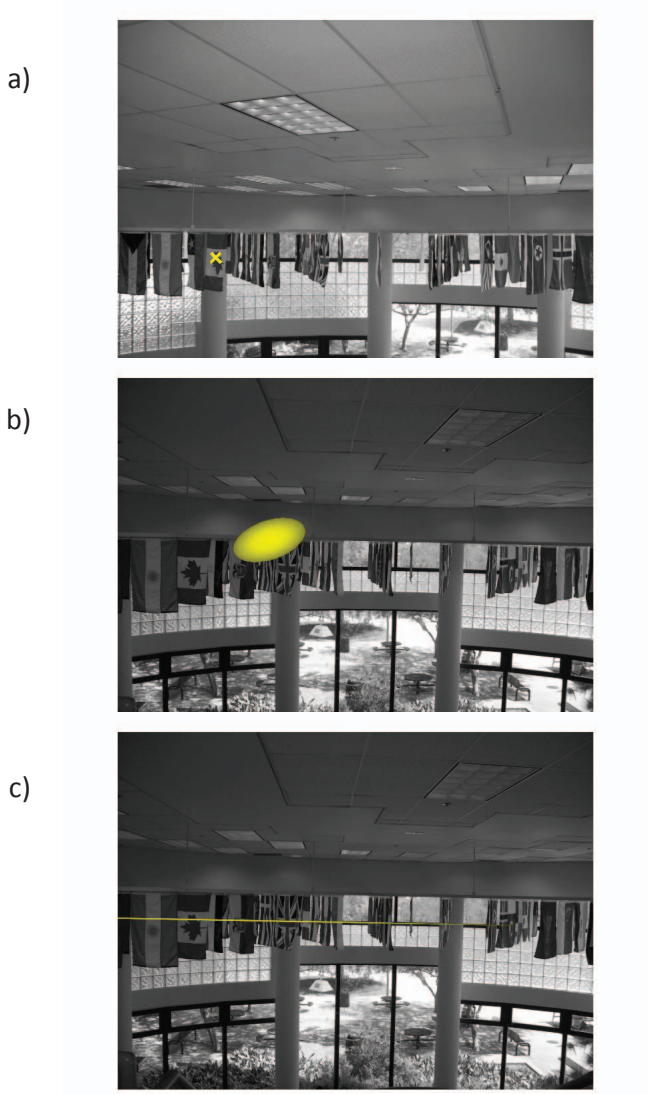


Figure 3.2: (a) Image with a point marked with a yellow '×' (b) High probability correspondence region over second image as captured in the JFD based on entire putative set (c) High probability correspondence region over second image found using JFD based on the best epipolar geometry found. Note that the ellipse in 'b' does not cover the corresponding point while that in 'c' it does. Also note that JFD for correct set of correspondence is narrower.

We can model the noisy mapping of the 2D features, u_{ij}^k, into the corresponding v_{ij}^k by the probability $p({\mathbf{v}_{ij}^k}^T|\mathbf{u}_{ij}^k)$. The form for this conditional probability is centered around the underlying, deterministic, 2D to 2D epipolar constraint where $\mathbf{F}$ is the fundamental matrix.

$$ {\mathbf{v}_{ij}^k}^T \mathbf{F}_{ij}(\theta) \mathbf{u}_{ij}^k = 0 \tag{3.5} $$

Let us draw a random sample of correspondences θ of size s from $\mathbf{X}_{ij}$ where $\theta = [\mathbf{x}_{d_1}, \cdots, \mathbf{x}_{d_s}]$ and $\mathbf{h}$ is the indices of the sampled match pairs. The above equation can be linearized by considering the tensor product of the corresponding points, $\mathbf{x}_{ij}^k = \mathbf{v}_{ij}^k \otimes \mathbf{u}_{ij}^k$, with dimension 9 by 1 and expressed in the form $\theta^T \mathbf{f} = 0$ where $\mathbf{f}$ is 9×1 version of $\mathbf{F}$. This linear form implies that the JFD models are Gaussian in the tensor space,

$$ p(\mathbf{x}_{ij}^k|\theta) \propto \exp - \left(\frac{\mathbf{L}_{ij}^k}{2} \right) \tag{3.6} $$

where the negative log-likelihood function, $\mathbf{L}_{ij}^k$, is given by

$$ \mathbf{L}_{ij}^k = {\mathbf{x}_{ij}^k}^T \mathbf{W}_{ij} \mathbf{x}_{ij}^k \tag{3.7} $$

where k varies from 1 to $|\mathbf{X}_{ij}|$.

Then we build their 9 by 9 homogeneous scatter matrix $\mathbf{V}_{ij} = \frac{1}{s}\theta\theta^T$ where θ is the 9 by s measurement matrix and i varies from 1 to n. This measurement matrix also appears in linear matching tensor estimation. Triggs [128] has found that the inverse of this matrix is a good estimate of the information tensor $\mathbf{W}_{ij} \approx \mathbf{V}_{ij}^{-1}$. In practice, we have to compute $\mathbf{W}_{ij} \approx (\mathbf{V}_{ij} + \text{diag}(\epsilon, \cdots, \epsilon, 0))^{-1}$ to regularize the inversion.

Thus, the JFD is parameterized by the homogeneous information tensor, $\mathbf{W}_{ij}$, which is symmetric positive definite 9 by 9 matrix generalizing the homogeneous information. We can estimate this from sample correspondences as follows.

The *conditional probability* of any match pair $(\mathbf{u}_{ij}^k, \mathbf{v}_{ij}^k)$ in $\mathbf{X}_{ij}$, given a set of correspondences θ , which is a sample of size s drawn from $\mathbf{X}_{ij}$ is given by the multivariate

Gaussian distribution function as follows

$$p(\mathbf{x}_{ij}^k|\mathbf{V}_{ij}(\theta)) = \\ |\frac{(\mathbf{V}_{ij}(\theta)+\epsilon)^{-1}}{(2\pi)^{4.5}}| \exp - \left(\frac{\mathbf{x}_{ij}^{k\,T}(\mathbf{V}_{ij}(\theta)+\epsilon)^{-1}\mathbf{x}_{ij}^k}{2\tau} \right) \tag{3.8}$$

where $\mathbf{V}_{ij}(\theta)$ is a 9 by 9 matrix constructed from θ for image pair $\{\mathcal{I}_i, \mathcal{I}_j\}$, as specified earlier. τ in the above equation is a scaling constant. In our experiments, the value of τ was set to 10^4. We will use this conditional probability function to sample from the correspondence space.

In the Figure 3.2, the physical implication of JFD is shown. The position of the JFD ellipse denotes accuracy of the correspondences and the eccentricity of the ellipse denotes the Gaussian inlier error in case JFD is formed only by inlier correspondences. Ideally, center of the JFD ellipse should be the corresponding point and the ellipse should be a straight line.

3.2.4 Fundamental Matrix Fitting Quality

Fundamental matrix fitting quality is a measure that evaluates how well the fundamental matrix $\mathbf{F}(\theta)$ fits the putative correspondence set $\mathbf{X}_{ij}$. It is represented by $\mu(\mathbf{F}_{ij}(\theta)|\mathbf{X}_{ij})$.

A M-estimator finds the weighted mean square error of all the data points for each candidate fundamental matrix. The candidate fundamental matrix for which the weighted mean square error is minimum is chosen as the 'best so far' in each sampling iteration. M-estimators look for maximum likelihood, and thus, M-estimators are Maximum likelihood estimators. Different M-kernels or weight functions give different M-estimators. We choose Welsh's [132] weight function as our M-kernel in this work. The negative exponential in the Welsh's weight function suppresses the effect of outliers on the evaluation of the quality of the fundamental matrix. The sum of such exponentials gives an M-estimate of the

fundamental matrix fitting quality. Our fundamental matrix fitting quality is given by

$$\mu(\mathbf{F}_{ij}(\theta)|\mathbf{X}_{ij}) = \sum_{k=1}^{|\mathbf{X}_{ij}|} \exp\left(\frac{-\delta^2(\mathbf{F}_{ij}(\theta)|x_{ij}^k).\sigma}{2}\right) \tag{3.9}$$

In our experiments, we have fixed $\sigma = 10^4$. It must be noted here that we do not multiply probability measures obtained from residual errors as in MAPSAC due to two reasons. Firstly, we assume conditional dependence unlike MAPSAC, which assumes conditional independence of correspondences. Secondly, multiplicative cost function would primarily be determined by the low probabilities associated with the outliers, which are large in numbers. Additive cost function would instead allow for suppressed fitting values of outlier correspondences without getting effected by them. We denote $\delta^2(\mathbf{F}(\theta)|\mathbf{x}_{ij}^k) = \delta_k^2$ as the Sampson's distance, which is given by

$$\delta_{ij}^{k\,2} = \frac{({v_{ij}^k}^T \mathbf{F}_{ij} u_{ij}^k)^2}{(\mathbf{F}_{ij}u_{ij}^k)_1^2 + (\mathbf{F}_{ij}u_{ij}^k)_2^2 + ({\mathbf{F}_{ij}}^T v_{ij}^k)_1^2 + ({\mathbf{F}_{ij}}^T v_{ij}^k)_2^2} \tag{3.10}$$

For notational simplicity, we denote $\mathbf{F}(\theta)$ as $\mathbf{F}$ in Equation3.10. $(\mathbf{F}u_{ij}^k)_1$ denotes the 1st entry of the vector $(\mathbf{F}u_{ij}^k)$ and so on. The expression captures the geometric error, where the denominator is the product of partial differentials of ${v_{ij}^k}^T \mathbf{F} u_{ij}^k$ and its transpose and the numerator is the square of ${v_{ij}^k}^T \mathbf{F}_{ij} u_{ij}^k$.

3.2.5 Degeneracy Measure

Figure 3.3 shows two image pairs. Image 'a' shows an inlier correspondel rejected as degenerate by our algorithm. Image 'b' shows the same image pair respectively with inlier correspondel that is accepted as non-degenerate by our algorithm. It can be easily noticed that the degenerate correspondel consists of correspondences that are have similar orientation and position with respect to each other. On the other hand, the non-degenerate correspondel consists of correspondences that have different orientation and position with

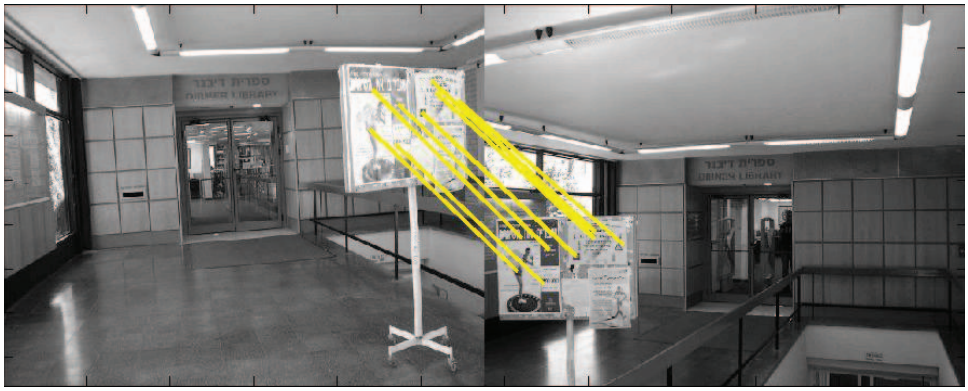
respect to each other. In other words, the scatter of the correspondences in the non-degenerate case is more. In this sub-section, we introduce our notion of degeneracy and its relation with associated scatter.

We try to avoid degeneracy by using a novel technique. The motivation is that each individual correspondence in the sample θ must add at least a small dimension to the motion models. We evaluate the mutual scatter of the correspondence pair $\mathbf{x}_{h_i}$ and $\mathbf{x}_{h_j}$ in the sample θ by considering the space induced by the matrix $[\mathbf{x}_{h_i}^T; \mathbf{x}_{h_j}^T]$, as captured by its two singular values. A zero for the lower singular value would suggest that the two correspondences are the same, i.e. they do not reduce the uncertainty [124, 136]in the estimate of the fundamental matrix. Two correspondences should reduce the uncertainty in the fundamental matrix estimation (using the linear method) by 2 dimensions. Of course, if we have 8 correspondences, then we have unique estimate of the matrix. The ratio of the low to the high singular value is a measure of the degeneracy of the two correspondences. The lower is the value of this ratio, more similar is the correspondences. A low ratio would suggest that the information in both the correspondences can be represented by either of them, so we would not need both. We weed out the motion models that are generated by such correspondences. The threshold β for this purpose should be very small, because although we do not want degenerate solutions, we are looking for the s-tuple of inlier correspondences that is most consistent with the putative set. Inliers in the set are similar to each other in some respect. Our threshold should be low such that it does not interfere in our search of inliers, while making sure that a model formed using degenerate s-tuple is not accepted.

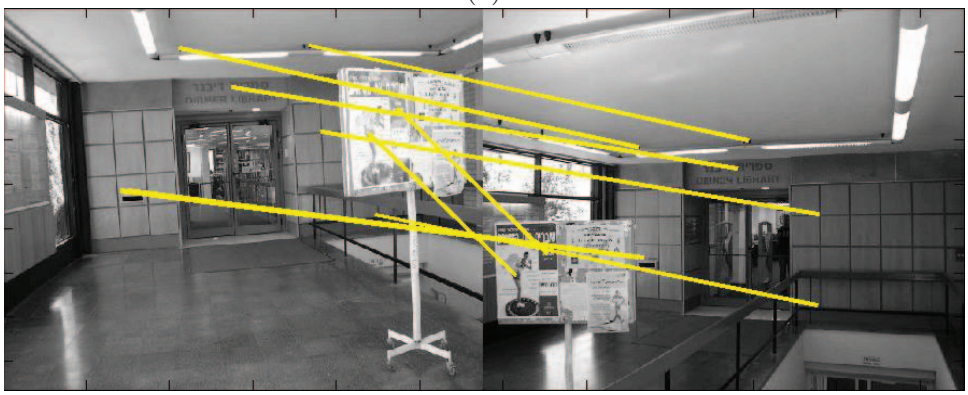
Our degeneracy measure is denoted by $\omega(\theta)$ where $\lambda_{1_{ij}}$ and $\lambda_{2_{ij}}$ are the singular values obtained by SVD. Note that the indices i and j stand for the correspondence between images and not images unlike earlier.

$$\omega(\theta) = \prod_{i=1}^{i=s-1} \prod_{j>i}^{j=s} \left(\frac{\lambda_{2_{ij}}}{\lambda_{1_{ij}}} > \beta \right) \tag{3.11}$$

If the value of the ratio is less than β the correspondence pair is degenerate and the value returned by this comparison is 0. $\omega(\theta)$ is obtained by the product of all such 0/1 decisions for all possible pairs in θ.



(a)



(b)

Figure 3.3: (a) A minimal set of correspondences that is considered to be degenerate (b) A non-degenerate minimal set.

3.2.6 Acceptance of a Sample

The photometric weights for each putative correspondence $p(\mathbf{x}_{ij}^k)$ can be obtained as mentioned in Sec. 3.2.2. The photometric proposal distribution for global hop is given by

$$p(\mathbf{x}_{ij}^k|\rho) = \frac{w_{ij}^k}{\sum_{k=1}^{|\mathbf{X}_{ij}|} w_{ij}^k} \tag{3.12}$$

Our geometric proposal distribution for local diffusion is given by

$$p(\mathbf{x}_{ij}^k|\theta^*) = \frac{p(\mathbf{x}_{ij}^k|\mathbf{V}_{ij}(\theta))}{\sum_{k=1}^{|\mathbf{X}_{ij}|} p(\mathbf{x}_k|\mathbf{V}_{ij}(\theta^*))} \tag{3.13}$$

The ratio of objective function is given by $y(\theta_t)$ as

$$y(\theta^t) = \frac{\mu(\mathbf{F}_{ij}(\theta^t)|\mathbf{X}_{ij})\omega(\theta^t)}{\mu(\mathbf{F}_{ij}(\theta^*)|X_{ij})\omega(\theta^*)} \tag{3.14}$$

If $y(\theta_t) > 1$

$$\theta^* = \theta^t$$

θ^t is accepted as 'best so far' sample if $y(\theta^t)$ is greater than 1. θ^* is the optimal 8-tuple of correspondences found so far. The first s-tuple is sampled using $p(\mathbf{x}_k|\rho)$ and thereafter MCMC sampling is triggered. This is repeated over T_N number of iterations.

Algorithm 1 $[\mu(\mathbf{F}_{ij}(\theta^*)|\mathbf{X}_{ij}), \frac{\mu(\mathbf{F}_{ij}(\theta^*)|\mathbf{X}_{ij})}{|\mathbf{X}_{ij}|}, \theta^*] = \text{BLOGS}(\mathcal{I}_i, \mathcal{I}_j, \alpha, \beta, T_N)$

Extract point features from images $\mathcal{I}_i$ and $\mathcal{I}_j$
Compute putative correspondences $\mathbf{X}_{ij} = [\mathbf{x}_{ij}^1, \cdots, \mathbf{x}_{ij}^{|\mathbf{X}_{ij}|}]$
Compute photometric hop proposal $p(\mathbf{x}_{ij}^k|\rho)$
Draw a minimal set, θ^* from $p(\mathbf{x}_{ij}^k|\rho)$ s.t. $\omega(\theta^*) = 1$ using β
if $\theta^* =$ NULL **then**
  RETURN
end if
Compute $\mu(\mathbf{F}_{ij}(\theta^*)|\mathbf{X}_{ij})$
for $t = 1 \rightarrow T_N$ **do**
  if $\alpha \leq$ rand (0, 1) **then**
    Draw a sample, θ^t from hop proposal $p(\mathbf{x}_{ij}^k|\rho)$
  else
    Draw a sample, θ^t from diffusion proposal $p(\mathbf{x}_{ij}^k|\theta^*)$
  end if
  Compute $\mu(\mathbf{F}_{ij}(\theta^t)|\mathbf{X}_{ij})$ and $\omega(\theta^t)$
  $y(\theta_t) = \frac{\mu(\mathbf{F}_{ij}(\theta^t)|\mathbf{X}_{ij})\omega(\theta^t)}{\mu(\mathbf{F}_{ij}(\theta^*)|X_{ij})\omega(\theta^*)}$
  if $y(\theta^t) > 1$ **then**
    $\theta^* = \theta^t$
  end if
end for

3.3 Image Match Scores

When two images capturing a common scene undergo big geometric transformations with respect to each other, it is difficult to establish correspondences photometrically. Usually not all photometric correspondences are consistent with the dominant underlying geometric transformation model (epipolar geometry/homography). Thus, photometrically generated correspondences are called putative correspondences and the correspondences consistent with the underlying dominant geometric transformation model are called inliers (see Figure 3.5) and the rest are called outliers. We observed that with the increase in transformation, the outlier rate in the photometrically established putative correspondences also increases. Also, the putative correspondences established are low in number. See Figure 3.4. Thus, the outlier rate in the putative correspondence set must be indicative of the geometric transformation. Thus, we can arrive at a photometric measure of image dissimilarity indicative of the geometric transformation. In this work, we propose a correspondence based score for image similarity, which we call the Cumulative Correspondence Score (CCS). The CCS is a score of the cumulative strength of putative correspondences which can be used to predict the outlier rate or image dissimilarity or geometric transformation.

3.3.1 Photometric Approximation to Geometric Match Score

The geometric match score is given by the inlier rate produces by the BLOGS algorithm. Now, let us move on to the problem in more details. The first problem that we face is of reliably estimating epipolar geometry and a geometric match score between widely separated images. Epipolar geometry estimations are computationally expensive and it is infeasible to compute them for every pair of images in a dataset. Also, many pairs of images might not be visually connected and thus it would lead to wastage of computation power if epipolar geometry is computed for all pairs of images in a dataset. In other words, we need to be selective about the image pairs that we want to consider for epipolar geometry

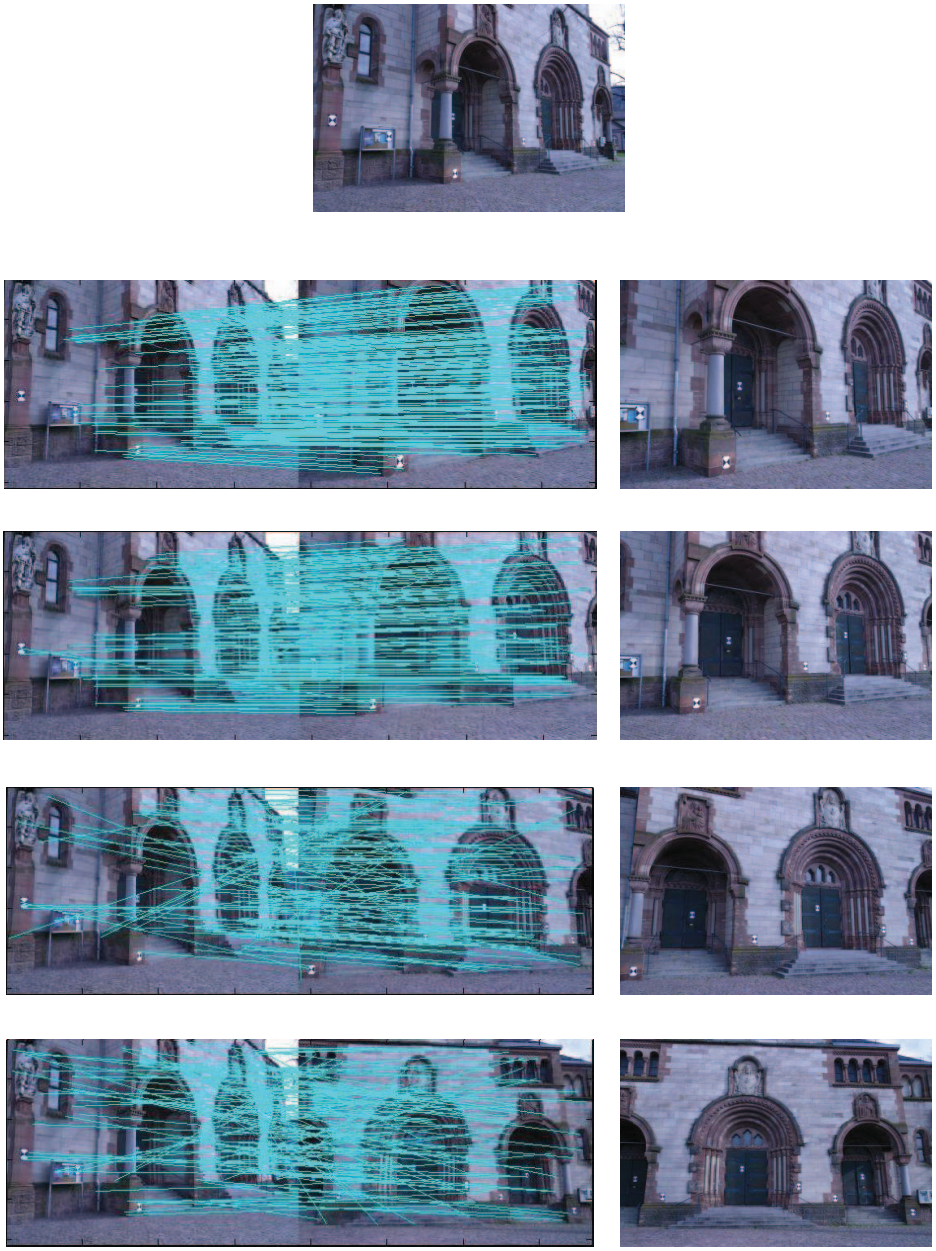


Figure 3.4: Putative correspondences in a sequence of 5 images. The first images are progressively matched with the next 4 images. Note that the inlier rate in the putative correspondence set between image pairs decrease with increase in geometric transformations between the pair of cameras.

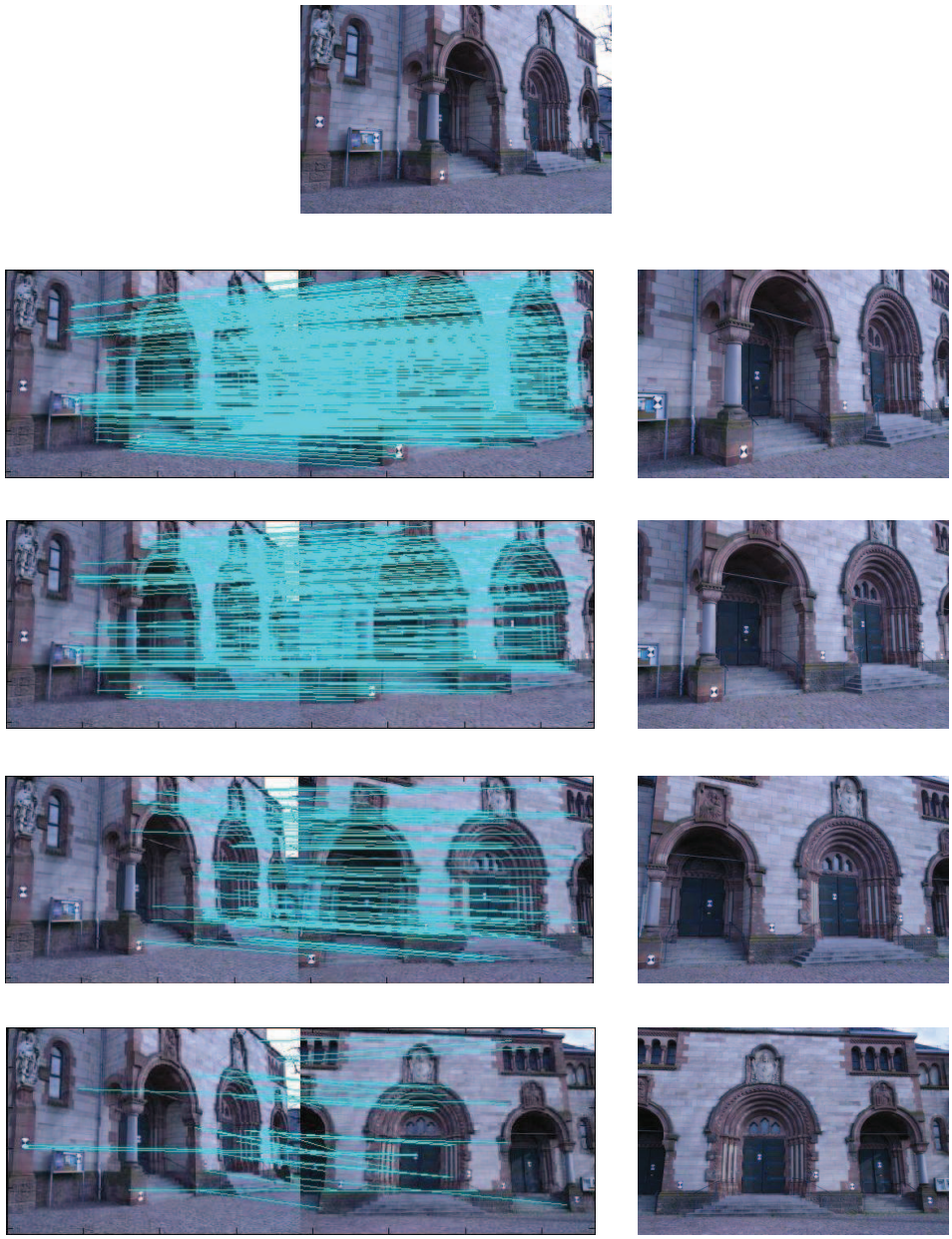


Figure 3.5: Correspondences inlier to the dominant fundamental matrix in a sequence of 5 images. The first images are progressively matched with the next 4 images. Note that the number of inliers between image pairs decrease with increase in geometric transformations between the pair of cameras.

estimation. Thus, the next problem is of estimation of a cheap indicative measure of match between all image pairs in a dataset to be used for the selection process. In other words, we need a cheap and close approximation of the geometric score to be able to waste lesser number of geometric estimations between images that do not match. Another part of this problem is to use this approximation to minimize the number of geometric estimations without sacrificing the graph connectivity.

Photometric similarity between two images $\mathcal{I}_i$ and $\mathcal{I}_j$ is defined based on the similarity between the sets of the point features found in these images. We used the SIFT [77] features in this work, but other features could also work. Let the features found in $\mathcal{I}_i$ and $\mathcal{I}_j$ be denoted by the sets $\{f_i^1, \cdots, f_i^{n_i}\}$ and $\{f_j^1, \cdots, f_j^{n_j}\}$, respectively. We take the similarity of a correspondence to be the reciprocal of the distance between SIFT features. This similarity is stored in n_i by n_j photometric feature similarity matrices $\mathbf{S}_{ij}$ for matches between each feature of $\mathcal{I}_i$ with each feature of $\mathcal{I}_j$. The accepted putative correspondences should have maximum similarity in both their rows and columns denoting the best mutual match. The match is as good as the similarity between features and its unlikeliness of having a better match with other features.

Let the similarity of the k-th such putative match be denoted by ρ_{ij}^k. And, ρ_{ij}^{kr} and ρ_{ij}^{kc} be the second highest values in the row and column of this match, respectively. The confidence in putative correspondences could be related to how different these second maximums are from the similarity of the putative correspondences, which is the maximum along the corresponding row and the column. The more is the difference; more is the confidence in the match being correct. We use the following combination to result in a normalized similarity for the k-th putative match.

$$w_{ij}^k = (1 - \exp^{-\rho_{ij}^k})^2 (1 - \frac{\rho_{ij}^{k_r}}{\rho_{ij}^k})(1 - \frac{\rho_{ij}^{k_c}}{\rho_{ij}^k}) \tag{3.15}$$

The combination results in high values for putative matches with high similarities *and* those for which the next best matches have low similarities. This photometric weight is

same as that used in [24]. Using these weights over the $|X_{ij}|$ putative matches, Cumulative Correspondence Score is estimated and is given by

$$\varphi_{ij} = 1 - \exp\left(-\kappa \log|\mathbf{X}_{ij}| \frac{\sum_{k=1}^{|\mathbf{X}_{ij}|} w_{ij}^k}{|X_{ij}|}\right) \tag{3.16}$$

where the photometric similarity between two images are exponentially related to the average similarity over the putative matches, $\frac{\sum_{k=1}^{|X_{ij}|} w_{ij}^k}{|X_{ij}|}$, and the number of putative matches, the $\log|X_{ij}|$, term. This similarity increases with the increase in number of putative correspondences and their average similarity. The constant term κ is fixed based on training data. In the experiments, we present some results with alternative combination forms. The above form resulted in the best performance.

Thus, we see that while ϕ_{ij} depends on both $\mathbf{F}_{ij}$ and $\mathbf{X}_{ij}$, φ_{ij} depends only on $\mathbf{X}_{ij}$, and thus, it is computationally much cheaper as it does not involve epipolar geometry estimation.

Algorithm 2 $\varphi_{ij} = \text{CCS}(\mathcal{V})$

```
Subsample all images in V to {I_1^s, I_2^s, ··· I_N^s}
Extract point features from all sub-sampled images
for i = 1 → N − 1 do
  for j = i + 1 → N do
    Compute putative correspondences X_ij = [x_ij^1, ··· , x_ij^|X_ij|]
    Compute the photometric confidence on X, i.e., w_ij^k
    Compute φ_ij using w_ij^k
    φ_ji = φ_ij
  end for
end for
```

3.4 Experiments

In the experiments, we tested the performance of BLOGS and the performance of CCS in approximating the match scores produced by BLOGS. Codes for competing epipolar geometry algorithms MAPSAC, NAPSAC and BEEM were obtained from the websites of

the respective authors [5, 121]. The codes to draw epipolar lines in Figure 3.1, Figure 3.6, Figure 3.7, Figure 3.8, Figure 3.9, Figure 3.10, Figure 3.11, Figure 3.12 were obtained from [2]. Points are marked in one image and in the other image the epipolar line should pass through the same point if the epipolar geometry found is correct. Please note that the epipolar lines are of good quality as they exactly or almost pass through the corresponding point in the pair of the marked images.

3.4.1 Datasets

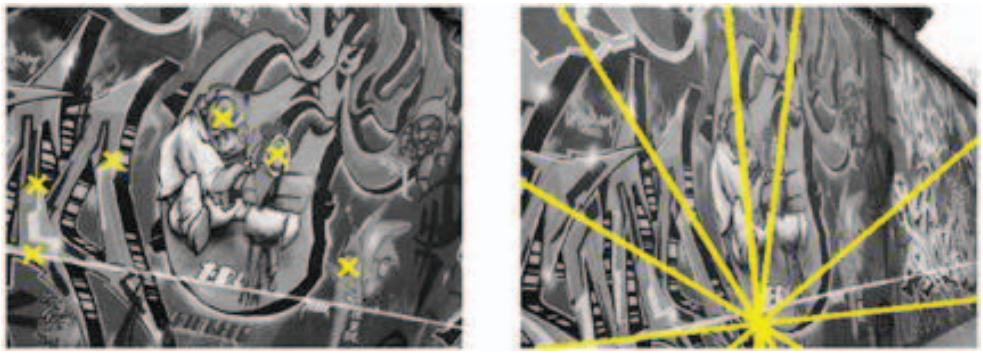
We have benchmarked performance of BLOGS and competing state-of-the-art algorithms on 20 image pairs. Ten of these image pairs are from our collection and ten other images have been collected from including ten images from other works. The images pairs in our collection in general have wider baseline. Such datasets have been addressed less in state-of-the-art research works. The Bluna image pair and KMsm image pair is from the WBS Image Matcher's website [10]. The Table image pair is from [7]. The Mountains and Path image pairs are from the 'Demoset' in the Nokia dataset [6]. The Books, Cactus, Library and Two Cars image pairs are from [5]. The Graphiti image pair is from [9]. The test data contains image pairs that have a very wide baseline, scale changes, rotation and occlusion. Such image pairs are not sufficiently addressed in the literature.

To find a good photometric approximation of the geometry based match score, tests are performed using the matching images in the Demoset of the Nokia [6] Lausanne dataset containing 105 images.

3.4.2 Ground Truth and Performance Evaluation

16 correspondences each were hand-marked in 20 wide-baseline images which served as a ground truth in order to estimate the error in fundamental matrix produced by our algorithm BLOGS for each pair of the 20 test images we have.

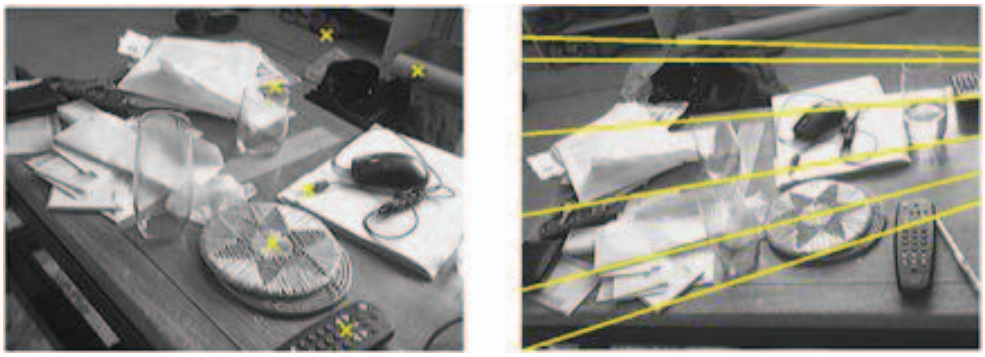
(1) Kmsm image pair



(2) Graphiti image pair

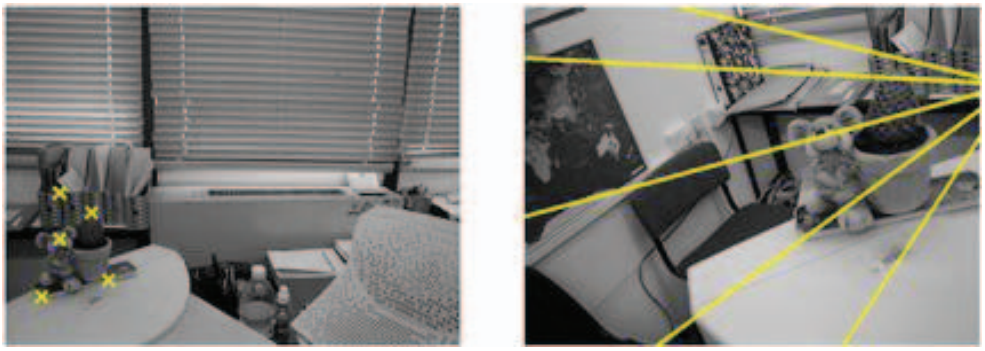
(3) Mountains image pair

Figure 3.6: Computed epipolar lines for *easy* image pairs. On the left image of each pair, we show marked points with '×'. The corresponding epipolar lines are drawn on the other image.



(4) Table image pair

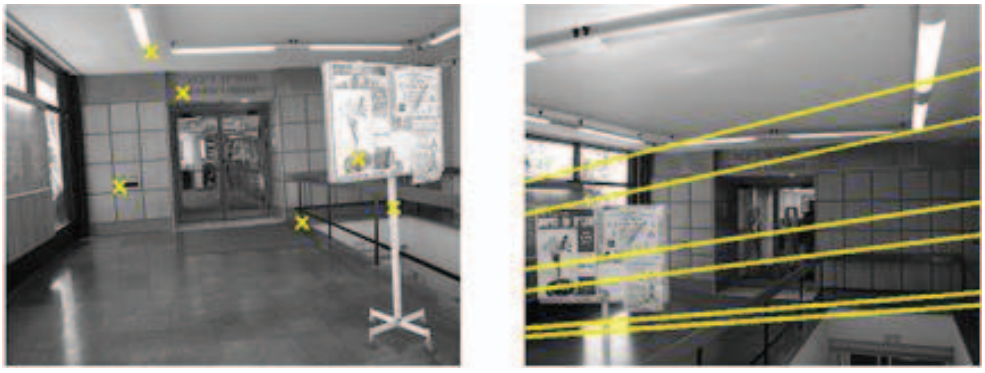
(5) Book image pair



(6) Cactus image pair

Figure 3.7: Computed epipolar lines for *easy* image pairs. On the left image of each pair, we show marked points with '×'. The corresponding epipolar lines are drawn on the other image.

(7) Two cars image pair

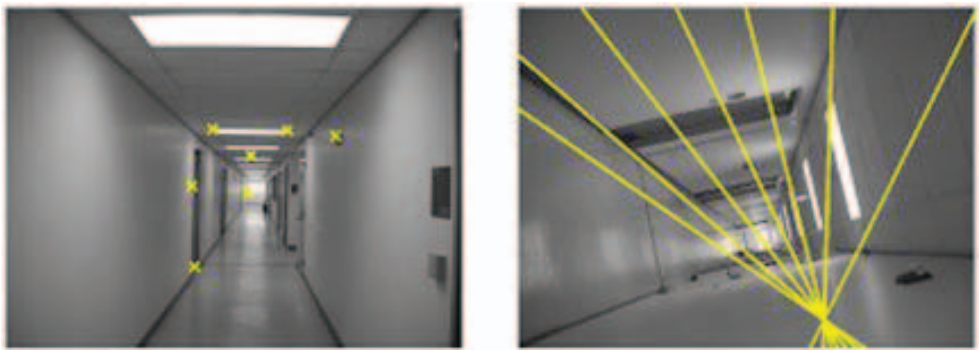


(8) Library image pair

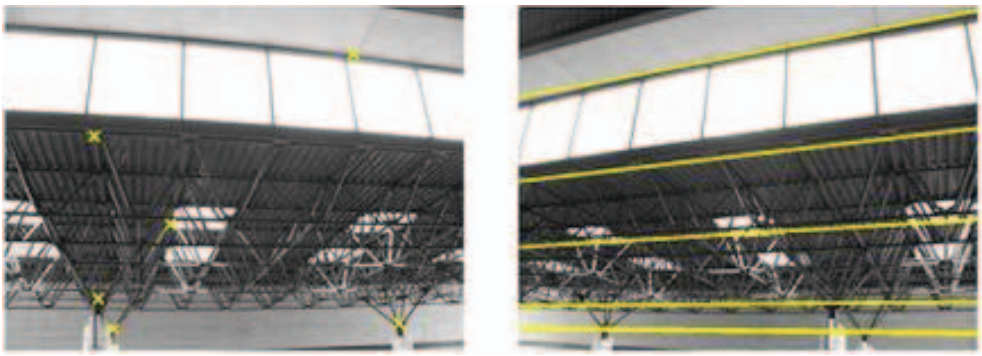
(9) Bluna image pair

Figure 3.8: Computed epipolar lines for *medium hard* image pairs. On the left image of each pair, we show marked points with '×'. The corresponding epipolar lines are drawn on the other image.

(10) Flags image pair



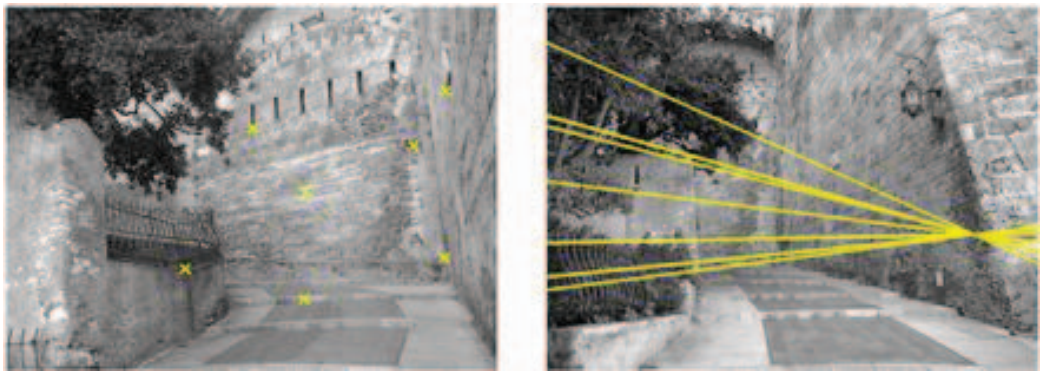
(11) Corridor image pair



(12) Steel mesh image pair

Figure 3.9: Computed epipolar lines for *medium hard* image pairs. On the left image of each pair, we show marked points with '×'. The corresponding epipolar lines are drawn on the other image.

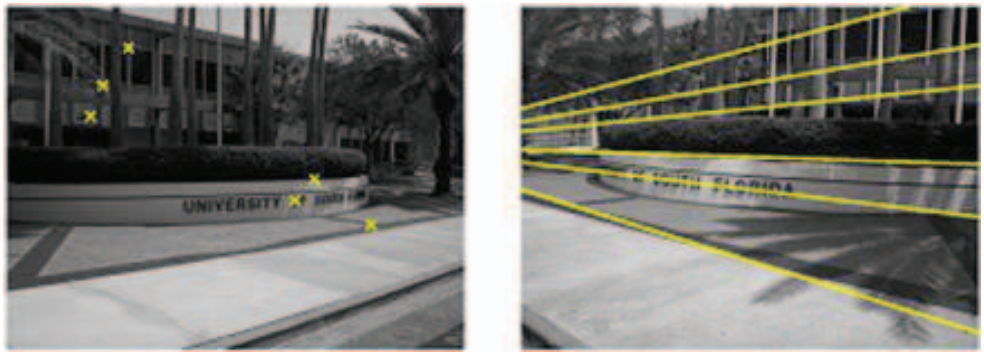
(13) Pillars image pair



(14) Path image pair

(15) Buiding image pair

Figure 3.10: Computed epipolar lines for *hard* image pairs. On the left image of each pair, we show marked points with '×'. The corresponding epipolar lines are drawn on the other image.



(16) University image pair

(17) Stones image pair

(18) Parking image pair

Figure 3.11: Computed epipolar lines for *hard* image pairs. On the left image of each pair, we show marked points with '×'. The corresponding epipolar lines are drawn on the other image.

(19) Cars image pair

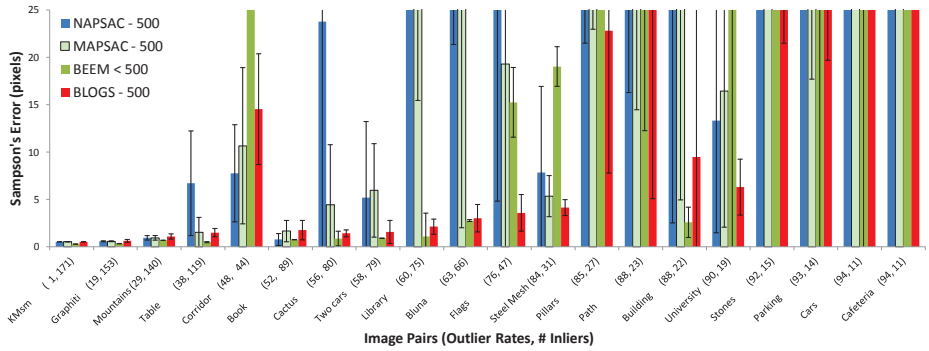
(20) Cafeteria image pair

Figure 3.12: Computed epipolar lines for *challenge* image pairs on which the proposed algorithm was tested. On the left image, we show the ground truth feature points with '×'. The corresponding epipolar lines are drawn on the other image.

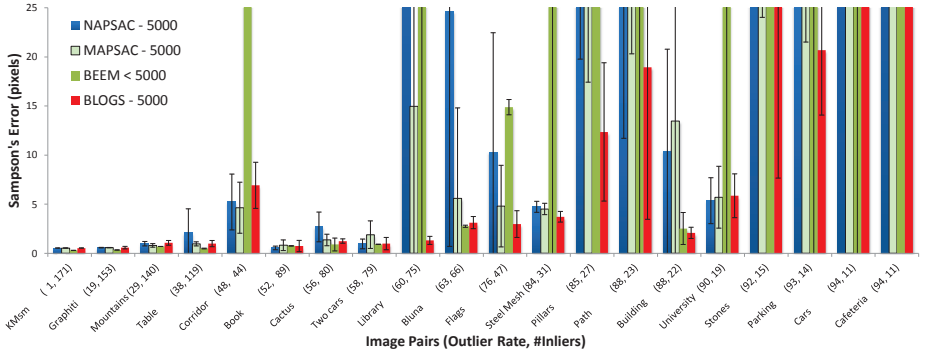
The Sampson's error is used to quantify the error in the fundamental matrix with respect to the 16 points hand-marked in the 20 images. Errors are averaged over 100 runs of our algorithm BLOGS and each of the competing algorithms NAPSAC, MAPSAC and BEEM with respect to the ground truth points marked. We also estimated the standard deviation of the errors.

We have manually quantified the number of inliers and the outlier rate in the putative correspondence set for each image pair to give an idea about the hardness of the epipolar geometry search required for each method. These are listed in increasing order of difficulty level given by the outlier rate in Figure 3.13.

For all matching images in the Demoset of the Nokia Dataset, the photometric match score given by Equation 3.17, Equation 3.18, Equation 3.19, Equation 3.20 and geometric match scores given by Equation 3.2.



(a)



(b)

Figure 3.13: Comparative performance of NAPSAC, MAPSAC, BEEM and BLOGS (our method). For each algorithm we list the mean of the root mean Sampson's distance for the 16 ground truth correspondences for separate 100 runs. The error bars are shown at one standard deviation. The horizontal axis represent the image pairs, sorted by their outlier rates (noted on the figure), which was manually determined. (a) Accuracies after 500 iterations. (b) Accuracies after 5000 iterations. The BEEM algorithm was allowed to exit early if its stopping condition was met.

3.4.3 Results

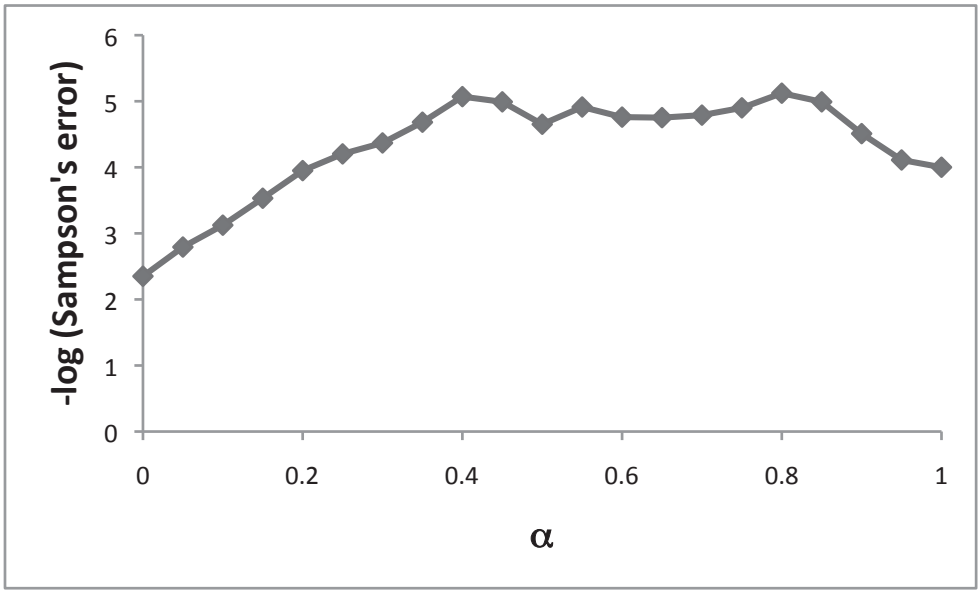
The quantified performances are shown in Figure 3.13. The outlier rate and number of inliers are mentioned with each image pair. MAPSAC performs better than both

NAPSAC and BEEM in most of the cases. We found that MAPSAC can handle high outlier rates as well. MAPSAC handling such high outlier rates is unreported to our knowledge. Our algorithm consistently performed better than NAPSAC and MAPSAC, while producing a little more pixel errors on one occasion. Very little differences might be ignored due to possible inaccuracy in hand-marked points, although points were marked with utmost care. Even on the hard images (the last two), our performance is better than the others. The results show that we attain almost the same accuracy in 500 iterations as MAPSAC attains in 5000 iterations and that BLOGS is capable of gaining more from barely sufficient number of inliers that might include degenerate inliers as well. Our algorithm is able to push to the limit of over 90 percent outliers with acceptable pixel errors in 5000 iterations in few cases. However, what value of pixel error is acceptable depends on the application. Note that the last two pairs in the figure are particularly challenging image pairs. Even BLOGS does not perform well on those images. They test the limits of current approaches, including ours, and help motivate future research to solve such problems. They motivate us to come up with even better solutions in future research.

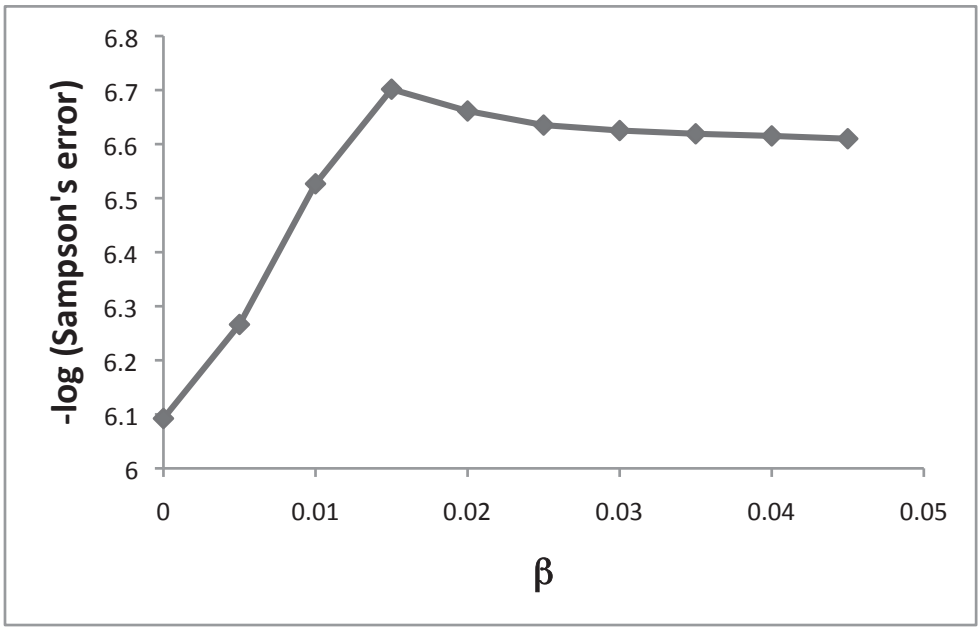
Table 3.1: Student's paired significance t-test associated probabilities between mean errors generated by 5000 and 500 iterations of BLOGS and 5000 iterations of MAPSAC, NAPSAC and iterations until convergence of BEEM not over 5000 iterations. Probability value less than 0.05 suggest that the hypothesis that the values are same can be rejected with more than 95% confidence; these are marked in bold.

Vs	BLOGS 5000	BLOGS 500
MAPSAC 5000	**0.004**	0.077
NAPSAC 5000	**0.014**	0.102
BEEM $<$ 5000	0.055	0.056

Figure 3.14a illustrates the essence of the mixing parameter and the degeneracy parameter. For each image pair, for each α value in the Figure 3.14a, the median of the error (root mean Sampson's distance from 16 hand-marked ground truth correspondences) over 100 executions was found. All median errors found for an image pair are normalized by dividing by the maximum median error found for that image pair across all α values



(a)



(b)

Figure 3.14: Variation of negative log of median of error (root mean Sampson's distance from 16 hand-marked ground truth correspondences) normalized for different choices of (a) the mixing parameter, α and (b) the degeneracy threshold parameter, β. The higher the value along the vertical axis, the better.

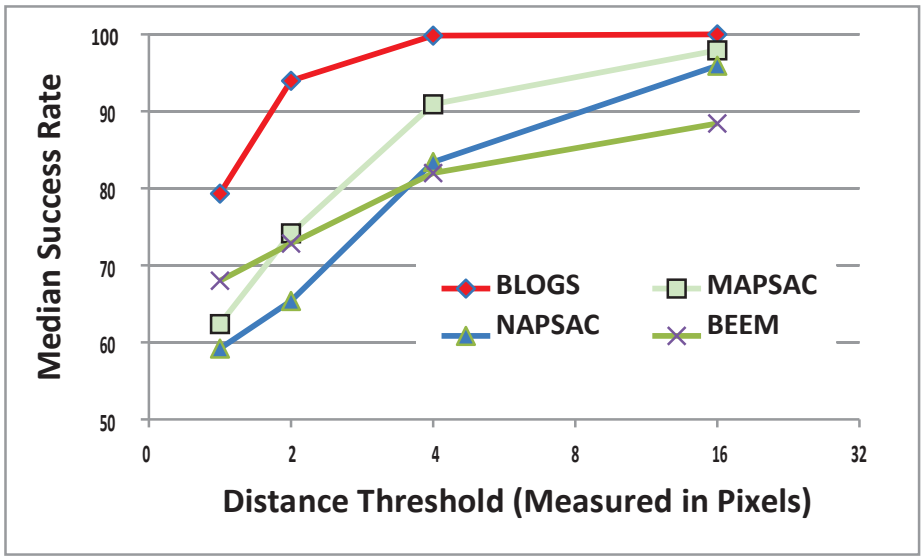


Figure 3.15: Comparative performance of MAPSAC, NAPSAC, BEEM and BLOGS based on median success rate across all image pairs in recognizing inliers within varying pixel thresholds.

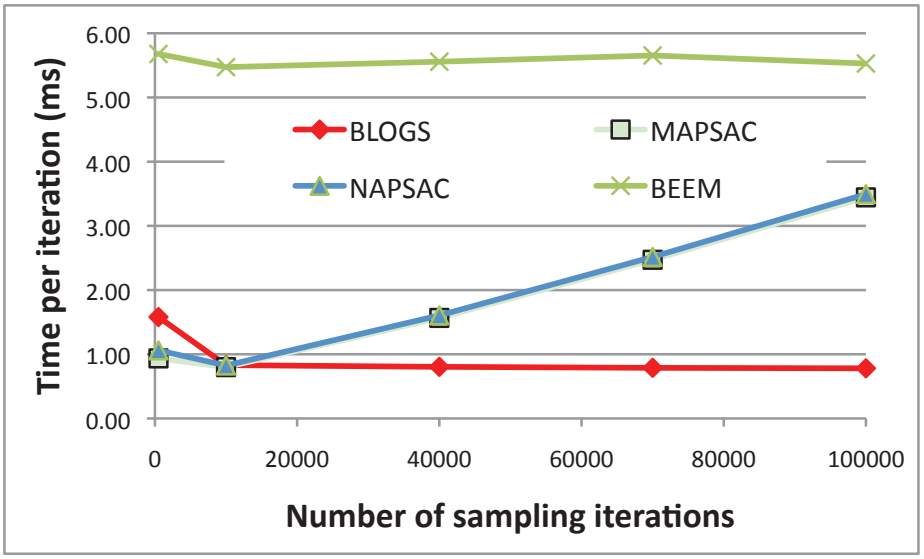


Figure 3.16: Per iteration wall-clock time in milliseconds taken by BLOGS, MAPSAC, NAPSAC and BEEM averaged over 500, 10000, 40000, 70000, 100000 iterations.

for which errors are evaluated. The normalization is done to limit the contribution of each image pairs in the cumulative error. The sum of the negative logarithm of normalized median of error for all images is plotted on the y-axis against α values in the x-axis. The higher the points along y-axis, the better they are. It can be seen that mixing is better than pure hop or pure diffusion, captured by $\alpha = 0$ and $\alpha = 1$. The square point giving best result in the figure corresponds to the case when diffusions were linearly increased with iterations. It can be seen that the α values around 0.5 produce the best results.

Figure 3.14b illustrates the essence of the degeneracy parameter. For each image pair, for each β value in the Figure 3.14b, the median of the error over 100 executions was evaluated and then the sum of negative logarithm of normalized median of error for all image pairs was plotted against each value of β. All median errors found for an image pair are normalized by dividing by the maximum median error found for that image pair across all β values for which errors are evaluated. It can be seen that the value of β chosen are small. From the plot it can be seen that initially the plot shows an increasing trend and then it goes down rapidly. This can be explained as follows. As long as the β value thresholds out only degenerate cases, the plot shows an increasing trend. The plot comes down rapidly after that. β values around 0.01 is thus suggested.

Tables 3.1 show the associated probabilities of Student's t-pair test on the mean Sampson's error for 100 executions of each algorithm on each image pair. In general, probability value less than 0.05 suggests that the hypothesis that the values are same can be rejected with more than 95% confidence and any value more than 0.05 suggests that statistically the values are the same. In Table 3.1, it can be concluded that the BLOGS in 5000 iterations gives mean errors significantly lesser than MAPSAC, NAPSAC each in 5000 iterations, while their errors in 5000 iterations are comparable to BLOGS in 500 iterations. This indicates that BLOGS produces results of similar quality in 10 times less number of iterations than MAPSAC and NAPSAC.

In Figure 3.15 we plot another performance measure. We compute the success rates, i.e. the percentage of inliers in the putative correspondence set that are within a

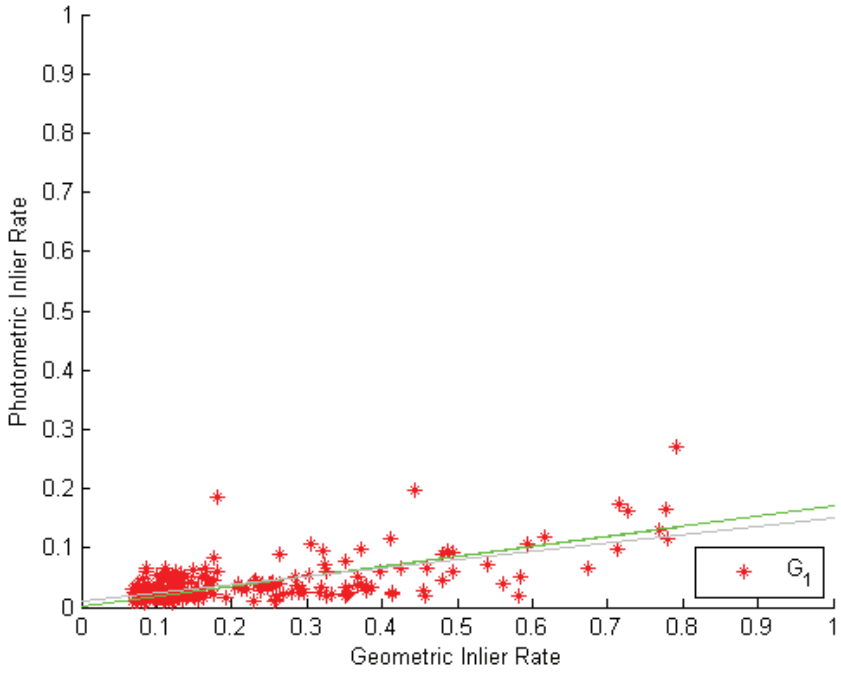


Figure 3.17: Correlation scatter plot with corresponding regression lines with (gray) and without (green) zero intercept between the photometric inlier rate produced by $\mathbf{G}_1$ and the geometric inlier rate. The correlation coefficient is 0.6509 without zero intercept. The slope of the zero-intercept line is 0.1706.

threshold distance of the computed epipolar geometry. The true inliers were identified manually. We report the median rate over all the images. While BEEM was allowed to meet its convergence, all other algorithms were executed up to 50000 iterations. We see that BLOGS is the best in terms of this measure too.

Figure 3.16 shows the wall-clock time per iterations for each algorithm plotted against increasing number of iterations. All codes were in MATLAB and executed on a Intel Dual Core processor @1.80 GHz with Windows7 Professional platform. BLOGS is the fastest. Stopping criterion of BEEM was disabled for this study. For BLOGS and BEEM, times per iteration are fairly constant, barring a slightly higher value for 500 iterations of BLOGS, which is due to the preprocessing overhead for the degeneracy estimation. This shows that for both BLOGS and BEEM, the time complexity is linear. However, BEEM

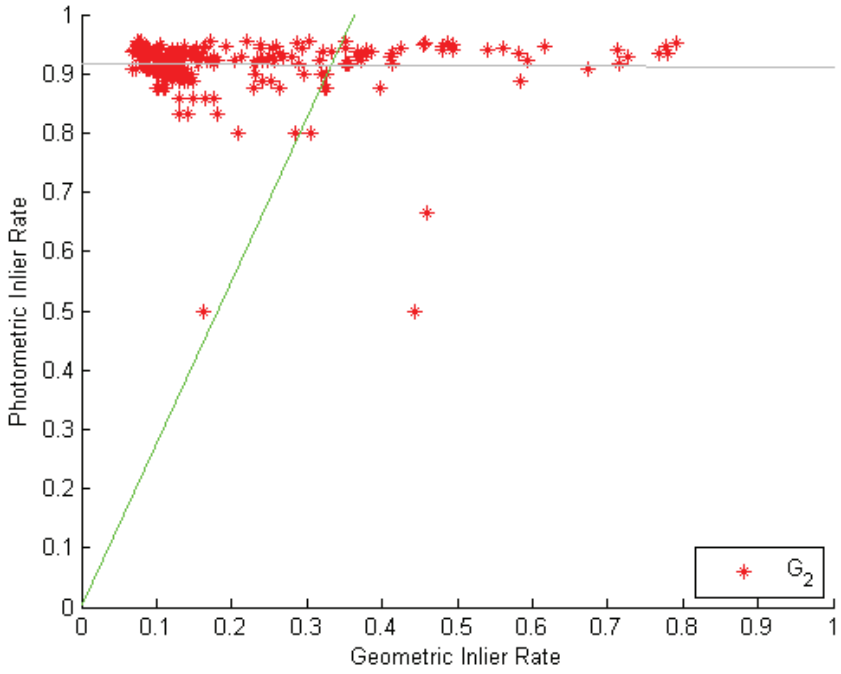


Figure 3.18: Correlation scatter plot with corresponding regression line with (gray) and without (green) zero intercept between the photometric inlier rate produced by $\mathbf{G}_2$ and the geometric inlier rate. The correlation coefficient is -0.0133 without zero intercept. The slope of the zero-intercept line is 2.7521.

takes about 5 times as much time per iteration as BLOGS. For MAPSAC and NAPSAC, the times per iteration are close to each other, but surprisingly increase with the number of iterations, suggesting that their time complexity is worse than linear.

$$\mathbf{G}_1(i,j) = 1 - \exp\left(-\frac{\sum_{k=1}^{|X_{ij}|} w_{ij}^k}{|X_{ij}|}\right) \tag{3.17}$$

$$\mathbf{G}_2(i,j) = 1 - \exp\left(-\log|X_{ij}|\right) \tag{3.18}$$

$$\mathbf{G}_3(i,j) = 1 - \exp\left(-12\frac{\sum_{k=1}^{|X_{ij}|} w_{ij}^k}{|X_{ij}|}\right) \tag{3.19}$$

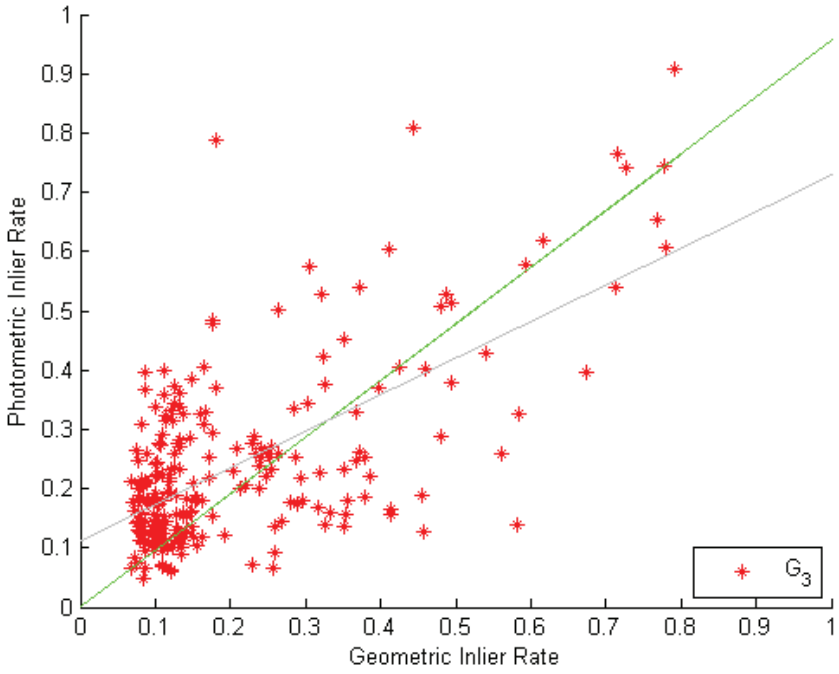


Figure 3.19: Correlation scatter plot with corresponding regression line with (grey) and without (green) zero intercept between the photometric inlier rate produced by $\mathbf{G}_3$ and the geometric inlier rate. The correlation coefficient is 0.6487 without zero intercept. The slope of the zero-intercept line is 0.9563.

$$\mathbf{G}_4(i,j) = 1 - \exp\left(-5\log|X_{ij}|\frac{\sum_{k=1}^{|X_{ij}|} w_{ij}^k}{|X_{ij}|}\right) \tag{3.20}$$

The four function in the Equation 3.17, Equation 3.18, Equation 3.19, Equation 3.20 mentioned are used to get an approximation of geometric inlier rate. The constants in the functions are dependent on the size of the images used. We found that the function of the form defined for CCS resulted in best correlation with the geometric inlier rate.

[t]

The figures 3.17, 3.18, 3.19, 3.20 respectively show the correlation plots between these these functions and ϕ_{ij}. We found that both Equations 3.17, Equations 3.19, Equation 3.20 are good. However, Equations 3.20 is slightly better in terms of correlation coefficient of the simple least squares linear regression line in gray and the slope of the zero

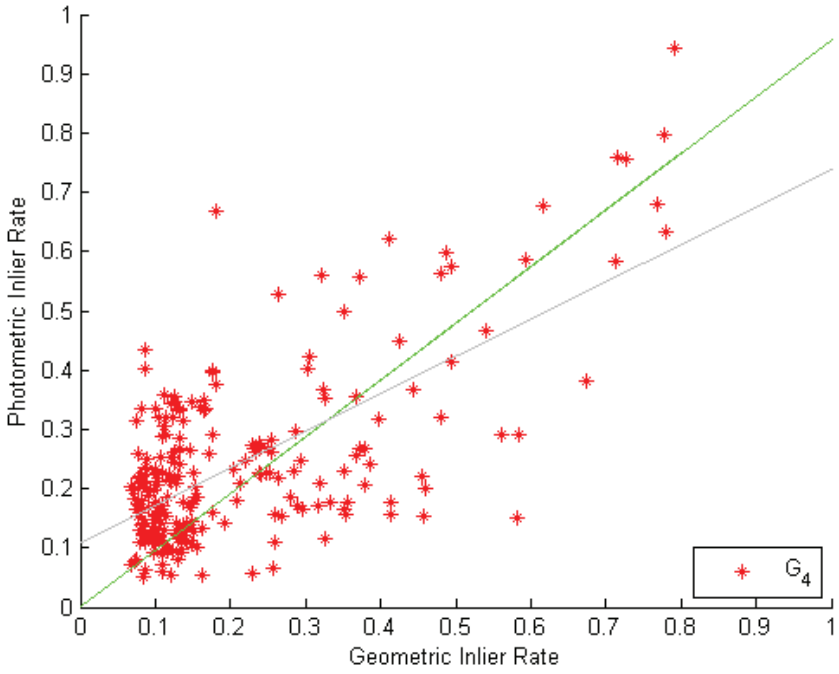


Figure 3.20: Correlation scatter plot with corresponding regression line with (grey) and without (green) zero intercept between the photometric inlier rate produced by $\mathbf{G}_4$ and the geometric inlier rate. The correlation coefficient is 0.6701 without zero intercept. The slope of the zero-intercept line is 0.9586.

intercept regression line in green thus used in this research. The null hypothesis significance test and more details are mentioned in Appendix C. Details on the zero-intercept regression line is mentioned in Appendix D.

Chapter 4 Basal Graph Structures and Geometry Based Image Organization

Image organization is an important problem and holds the key in many important applications. Geometry based applications like GPS and magnetometer tag noise detection and geometric walk-throughs are the focus of this dissertation. Thus, we propose basal graph structures for geometry based image organization. The term basal means minimal and foundational, minimal (basic) in terms of the algorithm and foundational (forms the base) in terms of the application.

We define our basal graphs as degree unconstrained or degree 2 constrained minimum spanning forests with all its edges satisfying a threshold and we refer degree unconstrained basal graphs as basal tree graphs and degree 2 constrained graphs as basal path graphs. Basal tree graphs are for planar image organizations without a loop and basal path graphs are for linear image organization as shown in Figure 1.4. Basal graphs are minimal and foundational graph structures for exploration and exploitation of information in a collection of images. Different types of image organizations are useful for different applications. For example, a collection of printed images are usually stored either side-by-side in an album in a tree arrangement on a plane or as a stack which are basically linear arrangements.

4.1 Background

Image organization [13, 26, 31, 50, 52, 68, 75, 106, 115] is done by looking at the relation between pairs of images and then deciding whether they are related or not. If they are related, graph structures are maintained to store the information about these relations. In geometry based image organization, ideally geometry based clustering would lead to the

most desirable result. However, quadratic cost of all pair geometry estimation is infeasible for large datasets. Thus, the objective is to identify all the relations in feasible time for large datasets. In the state-of-the-art, in order to meet this objective, photometric clustering is done, followed by seeking spanning trees to connect images and then expansion of the trees to form a denser graph.

4.1.1 Hybrid Approaches

Generally, only a fixed number of nearest neighbors [30] in exhaustive pair-wise estimates of appearance based similarity measure of images are put to test for geometrical verification to reduce the number of epipolar geometry estimations. In contrast, in this work we do not fix the number of nearest neighbors. Rather, our algorithm keeps exploring the neighbors until a stopping criterion is satisfied. Thus, we optimize both the connectivity and the number of geometric verifications simultaneously without a hard threshold on the number of neighbors. We use Cumulative Correspondence Score (CCS) as introduced in the previous chapter for photometric filtering.

4.2 Our Approach: CODIMSEG

We propose an algorithm COnnected Components Discovery by Minimally Specifying Expensive Graph (CODIMSEG) that uses an inexpensive approximate of an expensive complete graph to identify the connected components in the expensive graph by minimally specifying the expensive graph. The better the approximation is, the lesser would be the specification needed.

4.2.1 Problem Model, Notations and Mathematical Objective

Given a collection $\mathcal{V}$ of N images $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_N$, we want to efficiently discover the connectivity between them in the form of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{E}$ is the set of edges connecting them. Edge-connected pairs of images are those for which camera geometry can be faithfully estimated. We propose tree, $\boldsymbol{\Upsilon}$, and path, $\boldsymbol{\Pi}$, structures through the connected images. A denser graph, $\boldsymbol{\Lambda}$, of connected images can be efficiently discovered using the tree $\boldsymbol{\Upsilon}$. This graph $\boldsymbol{\Lambda}$ is then used for conversion of non-linear tree structures $\boldsymbol{\Upsilon}$ to linear path structures $\boldsymbol{\Pi}$. $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Pi}$ are called basal graph structures. The process of conversion of $\boldsymbol{\Upsilon}$ to $\boldsymbol{\Lambda}$ is called basal graph expansion. $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Lambda}$ might be split into c connected components if all the images are not connected. Thus, $\boldsymbol{\Upsilon} = \{\Upsilon_1, \cdots, \Upsilon_c\}$ and $\boldsymbol{\Lambda} = \{\Lambda_1, \cdots, \Lambda_c\}$. Further, each connected component might split into more number of paths in $\boldsymbol{\Pi}$ because it might not be possible to linearly represent a non-linear structure completely. Let the number of splits while finding path in each connected components be $s_1, \cdots, s_c$. Thus, $\boldsymbol{\Pi} = \{\Pi_{1,1}, \cdots, \Pi_{c,s_c}\}$.

We want to find the basal tree graphs that maximize the ratio of the number of edges in the basal tree graphs to the number of geometric estimations done. The numerator in the following equation is the match between the expensive graph and the basal tree graph, that is, the number of edges common in the two graphs and the denominator is the number of edges specified by an algorithm to produce the final basal tree graphs.

$$\Upsilon^{**} = \arg\max_{\Upsilon} \frac{|(\Upsilon \cdot \phi) > 0|}{|\phi^s > 0|} \tag{4.1}$$

The problem in the above optimization is that the full graph is never specified, because not fully specifying it is a part of the objective. Thus, we need an approximation of ϕ that we call φ and specify as less in ϕ as possible and attempt to maximally connect the graph using the specification. Thus, for graph connectivity, we look for a spanning forest.

4.2.2 Markov Chain Spanning Tree Diffusion for CODIMSEG

Ideally, the objective is to find a spanning forest with maximum number of edges in the expensive graph with geometric match scores as weights. However, we do not want to specify the expensive graph completely. We want to minimally specify the expensive graph without sacrificing the graph connectivity, such that the connected components or the number of connected components are not significantly different from the one achieved using pure geometric matching.

Let the low cost approximate of ϕ be denoted by φ.

$$\begin{aligned} \Upsilon_0 &= f_{MST}(\varphi_0) \\ [\varphi_b, \varepsilon] &= f_{REFINE}(\Upsilon_b, \phi^{\Upsilon_b}) \\ \Upsilon_{b+1} &= f_{MST}(\varphi_b) \end{aligned} \tag{4.2}$$

In the Equation 4.2, Υ_0 is the initial MST, Υ_b is the MST after b iterations and Υ_{b+1} is the MST after $b+1$ iterations in the Markov Chain process. The MST after $b+1$ iterations is dependent on φ_b updated by all refinements upto b iterations by estimation of ϕ over the MST edges at each iteration which we denote as Υ_b at iteration b. ε is the measure of improvement over the previous iteration in terms of the number of edges increased in Υ_b. The algorithm stops when $\varepsilon = 0$.

Our algorithm uses an inexpensive graph to minimally specify an expensive graph for basal tree graph discovery. We can use any photometric match score weighted graph. We use CCS φ to generate a complete graph weighted with photometric match scores. Next, we find a minimum spanning tree in this graph. For every pair of images I_i and I_j connected in this minimum spanning tree, we estimate the geometric match scores ϕ_{ij}^1 and ϕ_{ij}^2 and ϕ_{ij}^s is set to 1. If the thresholds Γ_1 and Γ_2 are not satisfied, the edge is set to a very high value ∞ in the photometric graph as an update. If the thresholds Γ_1 and Γ_2 are satisfied, then the edges are accepted in the basal tree graphs. Minimum spanning tree is estimated again

for the updated photometric graph and this process is repeated iteratively. If there is no change in the number of accepted edges, then the algorithm stops.

Figure 4.1 shows a graph visualization of image connectivity on the Lausanne Dataset [6] of 243 images after the CODIMSEG algorithm meets its stopping criterion. The red circles represent images with the image indices shown within the circle. We call these basal tree graphs. Note that many images were not connected. The image representations of the large basal tree graphs have been shown in Figure 4.6, Figure 4.7, Figure 4.8, Figure 4.9, Figure 4.10, Figure 4.11 and Figure 4.12 in the result section of this chapter. Note that all the connections are correct.

Algorithm 3 $[\Upsilon^*, \phi^1, \phi^2, opt] = \text{CODIMSEG}(\mathcal{V}, \phi^1, \phi^2, \alpha, \beta, T_N)$

$\varphi_{ij} = \text{CCS}(\mathcal{V})$
$b = 0$
$m = 1$
$\mathcal{G}'(\mathcal{V}, \mathcal{E} = \varphi)$
while m **do**
  $m = 0$
  $\Upsilon^t = \text{PRIM}(\mathcal{G}')$
  for $i = 1 \rightarrow N - 1$ **do**
    if $\phi^1_{\Upsilon^b(i,1),\Upsilon^b(i,2)} == 0$ **then**
      $[\phi^1_{\Upsilon^b(i,1),\Upsilon^b(i,2)}, \phi^2_{\Upsilon^b(i,1),\Upsilon^b(i,2)}] = \text{BLOGS}(I_{\Upsilon^b(i,1)}, I_{\Upsilon^b(i,2)}, \alpha, \beta, T_N)$
    end if
    if $\phi^1_{\Upsilon^b(i,1),\Upsilon^b(i,2)} > \Gamma_1$ AND $\phi^2_{\Upsilon^b(i,1),\Upsilon^b(i,2)} > \Gamma_2$ **then**
      $m = m + 1$
    else
      $\varphi_{\Upsilon^b(i,1),\Upsilon^b(i,2)} = \infty$
      $\varphi_{\Upsilon^b(i,2),\Upsilon^b(i,1)} = \infty$
    end if
  end for
  $opt(b) = m$
  $b = b + 1$
end while
$\Upsilon^* = \Upsilon^b$

Figure 4.2 shows how the number of connected edges increase using GIST and CCS. CCS starts with a lead, but for both CCS and GIST, CODIMSEG shows the same trend.

This indicates that while CCS is better than GIST, CODIMSEG can work with any kind of initialization. However, the expensive graph edges specification would be done more often.

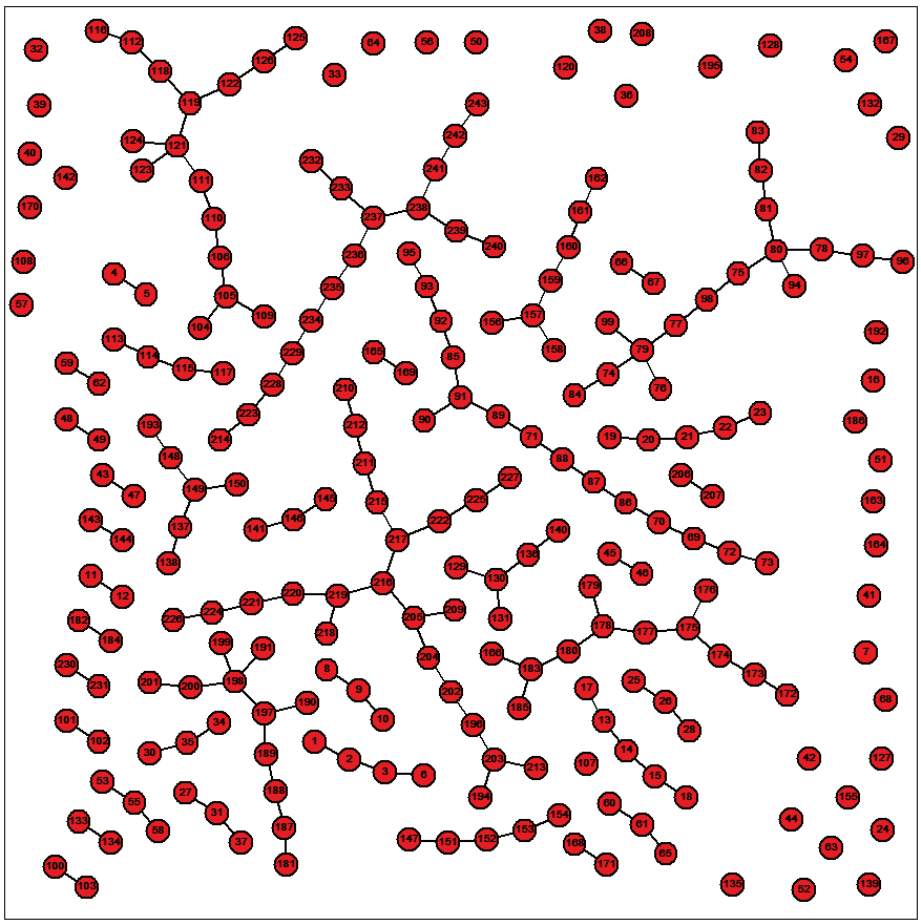
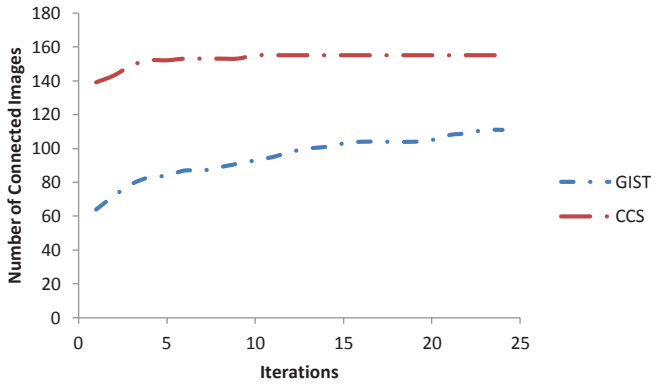
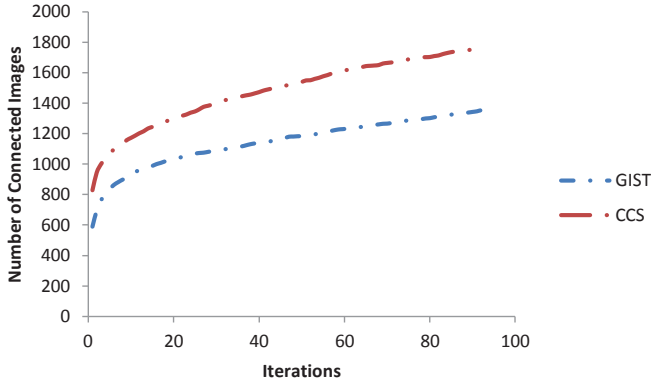


Figure 4.1: Basal tree graphs for Lausanne Dataset with 243 images represented as red circle with image indexes within them. The connected dots represent connected images. Notice that there are few big clusters and many images that are unconnected. This result was produces at a threshold such that no false positives or false edges are allowed at all. This typically also leads to over-clustering.

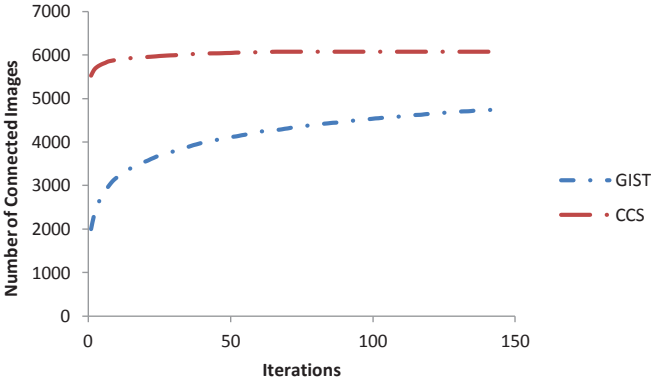
Figure 4.3 is a visualization of MST between geographically spread cameras. Using GPS and magnetometer tags, we were able to plot the camera position is the figure. Next, we show the initial MST and the final MST. Many edges in the initial MST were rejected. Geographically distant cameras might not have been used to capture images of a common



(a)



(b)



(c)

Figure 4.2: Increasing edges in the basal graphs with successive iterations of the CODIMSEG algorithm for a) Lausanne, b) Oxford and c) ArtQuad datasets. The figures show how the number of geometrically verified edges in the basal tree graphs increase with iterations. The algorithm stops when there is no increase in edges for consecutive iterations for given number of times

scene. Thus, we see that most often the cameras that are geographically distant are connected by edges that are rejected.

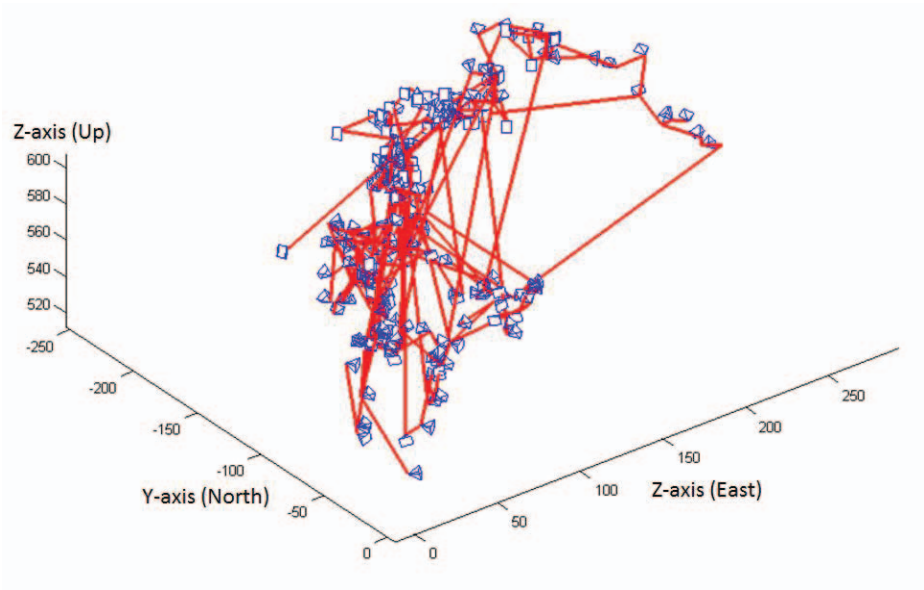
4.2.3 Basal Tree Graphs Expansion using CODIMSEG

Graph expansion is done to get a denser graph using the basal graphs. Graph expansion is done by looking at the transitive closures [13]. Transitive closures mean that if two vertices are not directly connected and if they are connected by path of length two in the graph, then verify if the two vertices are directly connected. This is done exhaustively in [13]. In our algorithm, we reuse the CODIMSEG algorithm to connect the vertices directly connected to a vertex using a basal tree graphs. Once this is done for all vertices, a second round of expansion is done, followed by third round and so on till the expansion leads to an extra edge.

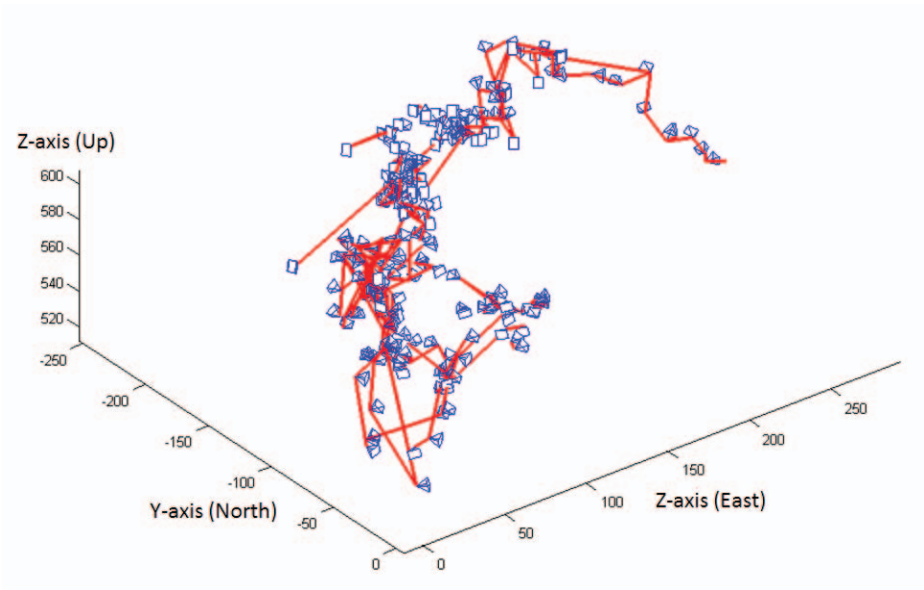
Algorithm 4 Λ = BASAL-TREE-GRAPH-EXPANSION($\Upsilon^*, \mathcal{V}, \phi^1, \phi^2, \alpha, \beta, T_N$)

```
Λ = Υ*
m = 1
while m do
  m = 0
  for i = 1 → N − 1 do
    For vertex I_i in Υ*, find the set of all vertices V_i reachable from I_i.
    [Υ', φ^1, φ^2, opt] = CODIMSEG(V_i, φ^1, φ^2, α, β, T_N)
    m = m + sum(opt)
    if sum(opt) > 0 then
      Λ = append(Λ, Υ')
    end if
  end for
end while
```

Figure 4.4 shows the graph in figure 4.1 after graph expansion. Note that many edges were added, however, the edges are added in the same connected component. Thus, connected component are still the same. The next phase is the to find basal path graphs.



(a)



(b)

Figure 4.3: Tree computation for the Nokia Challenge dataset. Each image is represented by a point at the corresponding GPS tagged locations, in local ENU (East North Up) coordinates, with negative Y-axis as north and positive X-axis as east. The reference camera at origin is encircled. We use the GPS tags just for visualization of the results. They are not used in the computation of the tree. (a) Initial photometry based minimum spanning tree, (b) Final tree structure.

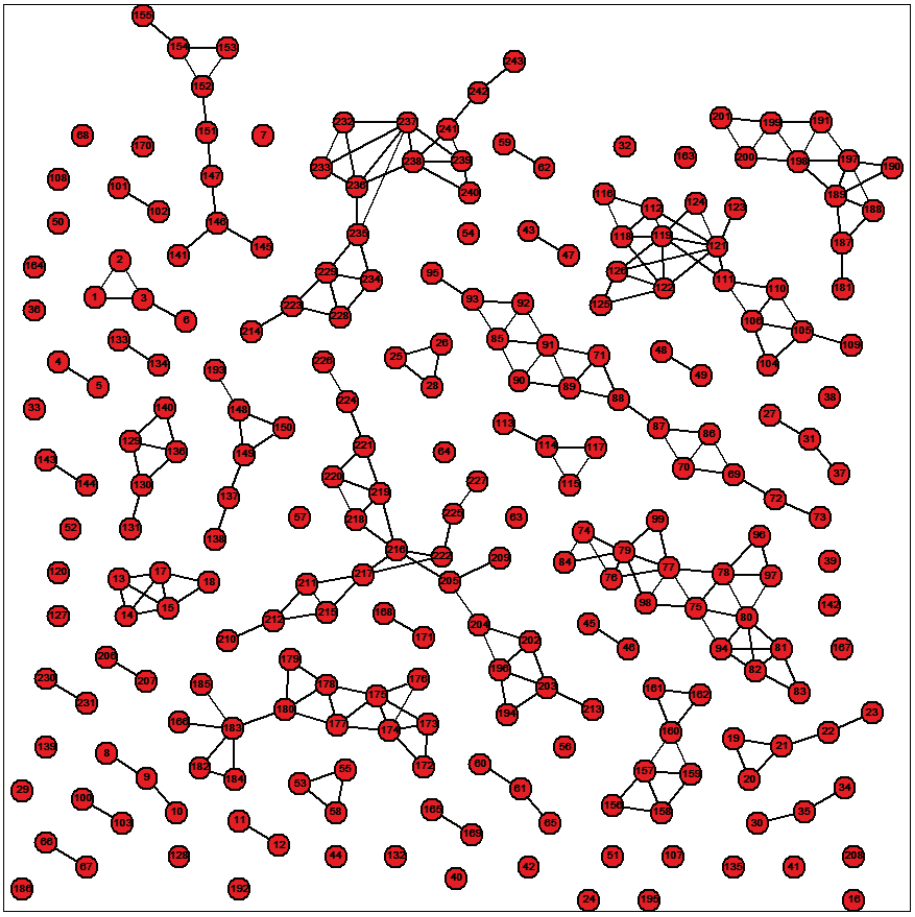


Figure 4.4: Expanded basal trees for the Lausanne Dataset. Notice the changes in the number of edges specially in the large clusters.

4.2.4 Basal Path Graphs using Minimum Hamiltonian Path Approximation Algorithms

After the basal graph expansion is done, we find the basal path graphs using a minimum Hamiltonian path approximation algorithm [16, 22, 102]. We choose two of these algorithms. One is the minimum spanning tree based approximation that has a theoretical bound on performance and the other is the best known heuristic for the problem, that is, the Lin-Kernighan algorithm. We use a version of the Lin-Kernighan algorithm called the chained Lin-Kernighan algorithm [1, 16].

Minimum spanning tree based approximation is 2-approximation algorithm and might not lead to a very good result, however, it just requires a tree for which the basal tree graphs are sufficient and the expanded basal graphs are not required. Lin-Kernighan algorithm can be used on any graphs if the edges not connect are set to very high weights indicating ∞. Thus, Lin-Kernighan algorithm can be applied to basal tree graphs as well as expanded basal graphs producing basal path graphs. However, it would lead to more connections in the basal path graphs if expanded basal graphs are used.

Algorithm 5 Π = GEOMETRIC-WALKTHROUGHS(Υ^*, Λ, $\mathcal{V}$, *choice*)

```
if choice == 0 then
   G(V, E = Υ*)
   Π = DFS(G) /* Revisit Nodes */
else
   G(V, E = Λ)
   Π = LIN-KERNIGHAN(G)
end if
```

Figure 4.5 shows a basal tree graph from Figure 4.1. The basal tree graph is then converted to basal path graph with and without graph expansion. The results show that the graph expansion is an important intermediate step before the non-linear tree structure is converted to a linear path structure.

Figure 4.6 to Figure 4.12 show various basal tree graphs with more than ten vertices. Figure 4.13 to Figure 4.17 show corresponding expanded graphs and basal path graphs.

4.2.4.1 Minimum Spanning Tree based Approximation

Minimum Spanning Tree based approximation of a Hamiltonian Path is done by using a Depth-First-Search algorithm on the graph. A Hamiltonian Path can be formed which would always be less than twice the length of the desired Hamiltonian cycle as it is a 2-approximation algorithm as each edge is traversed twice.

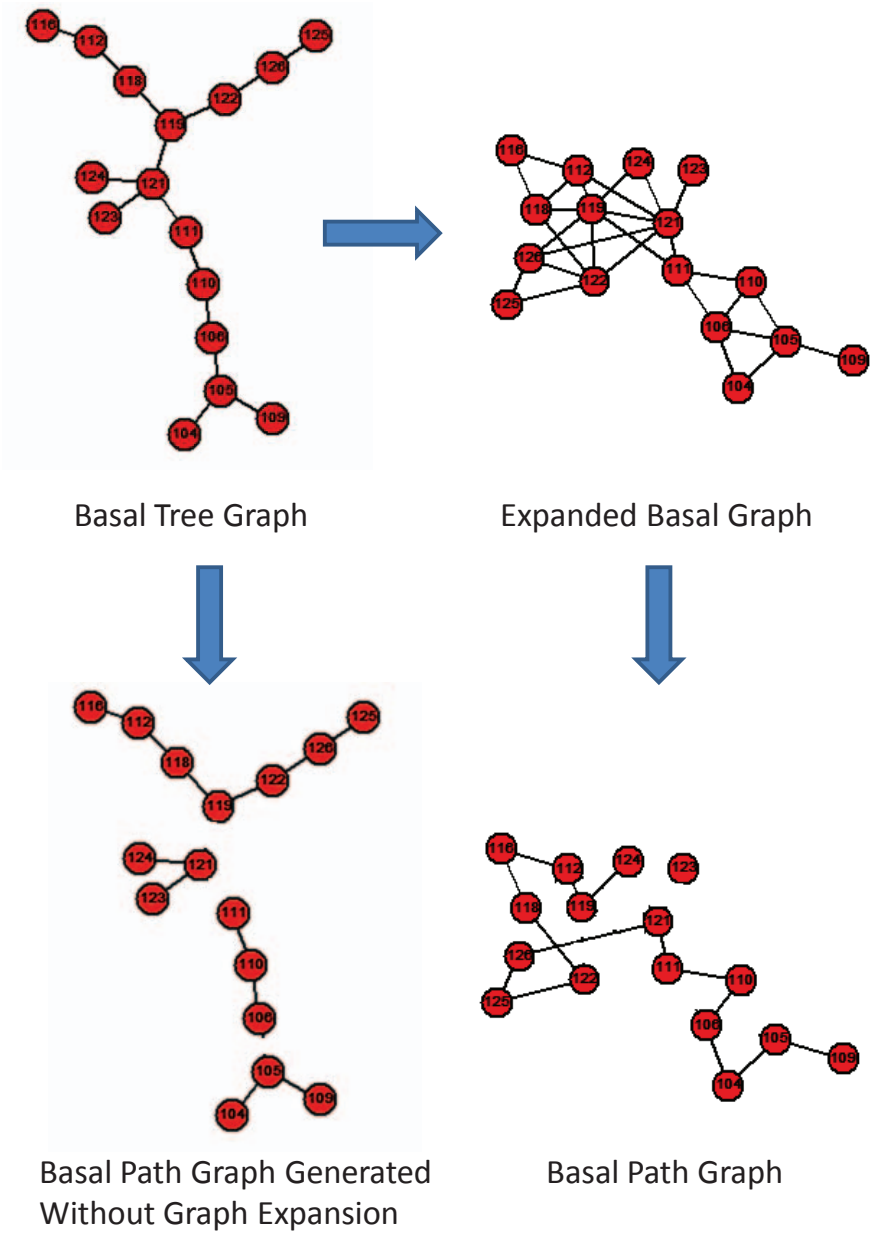


Figure 4.5: Expanded graph and basal paths in the basal tree shown in Figure 4.6. Notice that one image could not be accommodated in the path because it could not have two neighbors. Also notice that if expansion were not done, the basal path graph would have 4 components.

4.2.4.2 Chained Lin-Kernighan Heuristic

Lin-Kernighan algorithm is one of the best heuristics for the finding a Hamiltonian Path. It works by doing 2-opt, 3-opt, or upto 5-opt moves which involves exchanges of edges between 2, 3 or up to 5 edges to find a better path.

4.3 Experiments

We used different types of datasets to test various aspects of our algorithms in our research. Nokia Dataset is a challenging dataset that has never been responded in research so far except by us. Oxford building dataset has been used in CBIR applications and we use it to leverage geometric information in it. Oxford dataset has many images that do not match with any other image. Such images are called distractors and pose a challenge in image organization. Oxford building and the ArtQuad dataset tests the scalability of our algorithm. ArtQuad dataset has been used for large scale 3D modeling of a geographical region. Nokia dataset and ArtQuad dataset have GPS information. However, we do not use the GPS tag information for image organization. In the next chapter, GPS tags are verified using the organized images from this chapter. More details on the datasets are as follows. We used few graph related codes from [3, 8]. We used chained Lin-Kernighan heuristic from [1].

4.3.1 Datasets

The Nokia dataset [6] has two datasets in it: the 'Lausanne' dataset and the 'Demoset'. The Lausanne dataset has 243 images and the Demoset has 105 images collected using a commercial (Nokia 6210 Navigator) phone equipped with GPS and magnetometer sensors. Both the datasets are very challenging. The Lausanne dataset is a very well-collected dataset as it spans across a large geographical area and still visually connected,

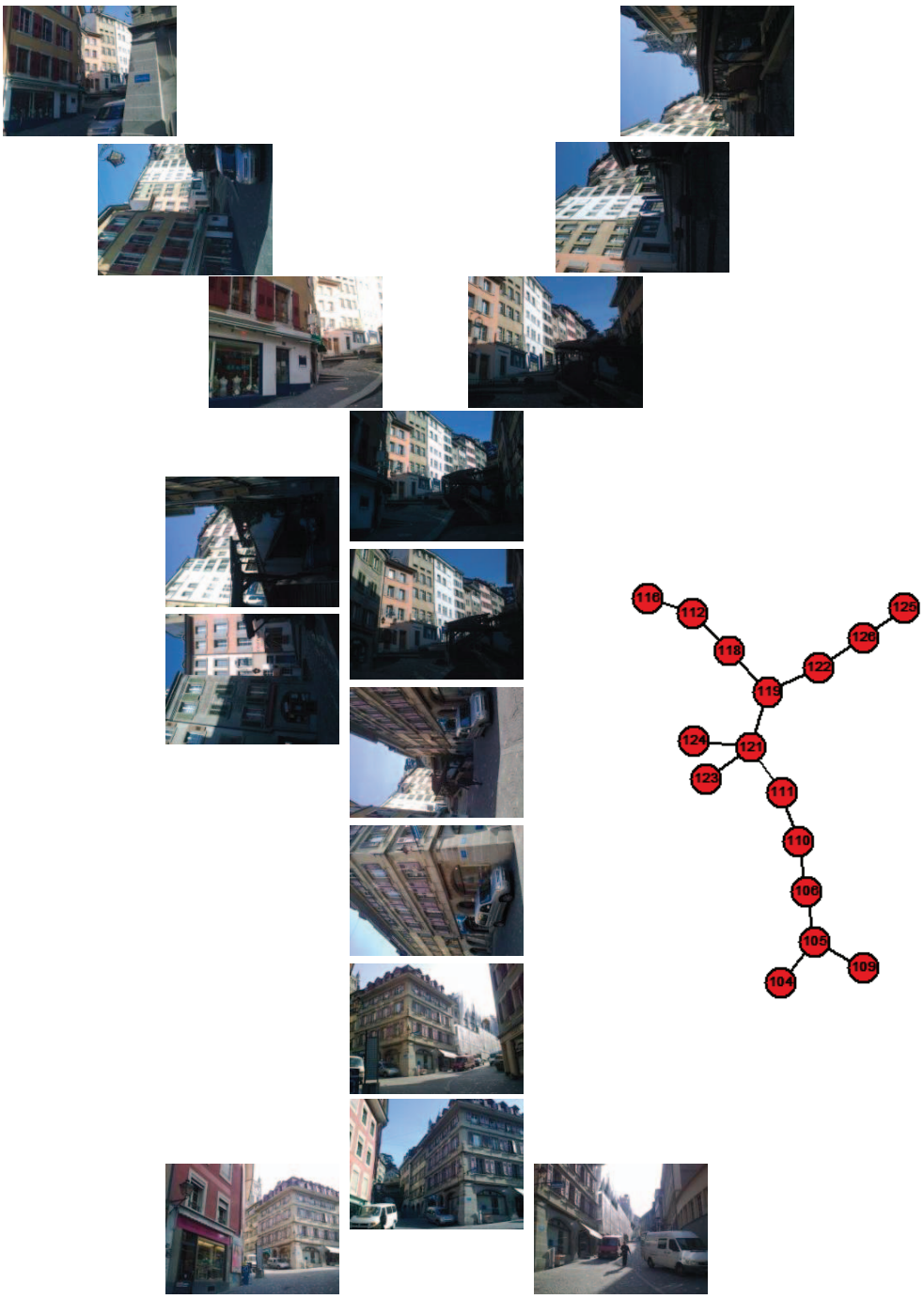


Figure 4.6: One of the basal trees for the Lausanne Dataset. Notice the large view-point changes in the images.

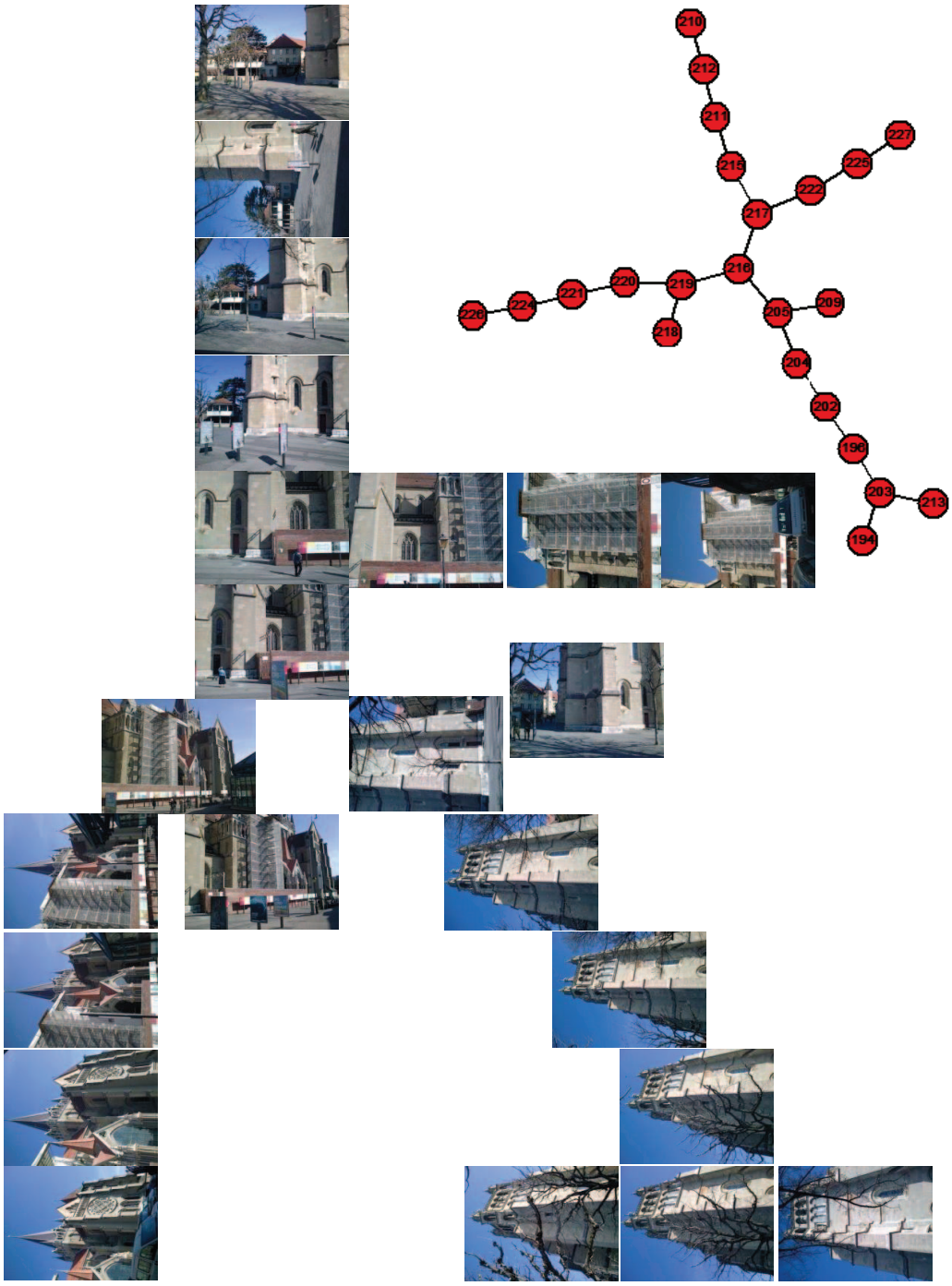


Figure 4.7: One of the basal trees for the Lausanne Dataset. Notice the significant difference between the leaf nodes of this tree.

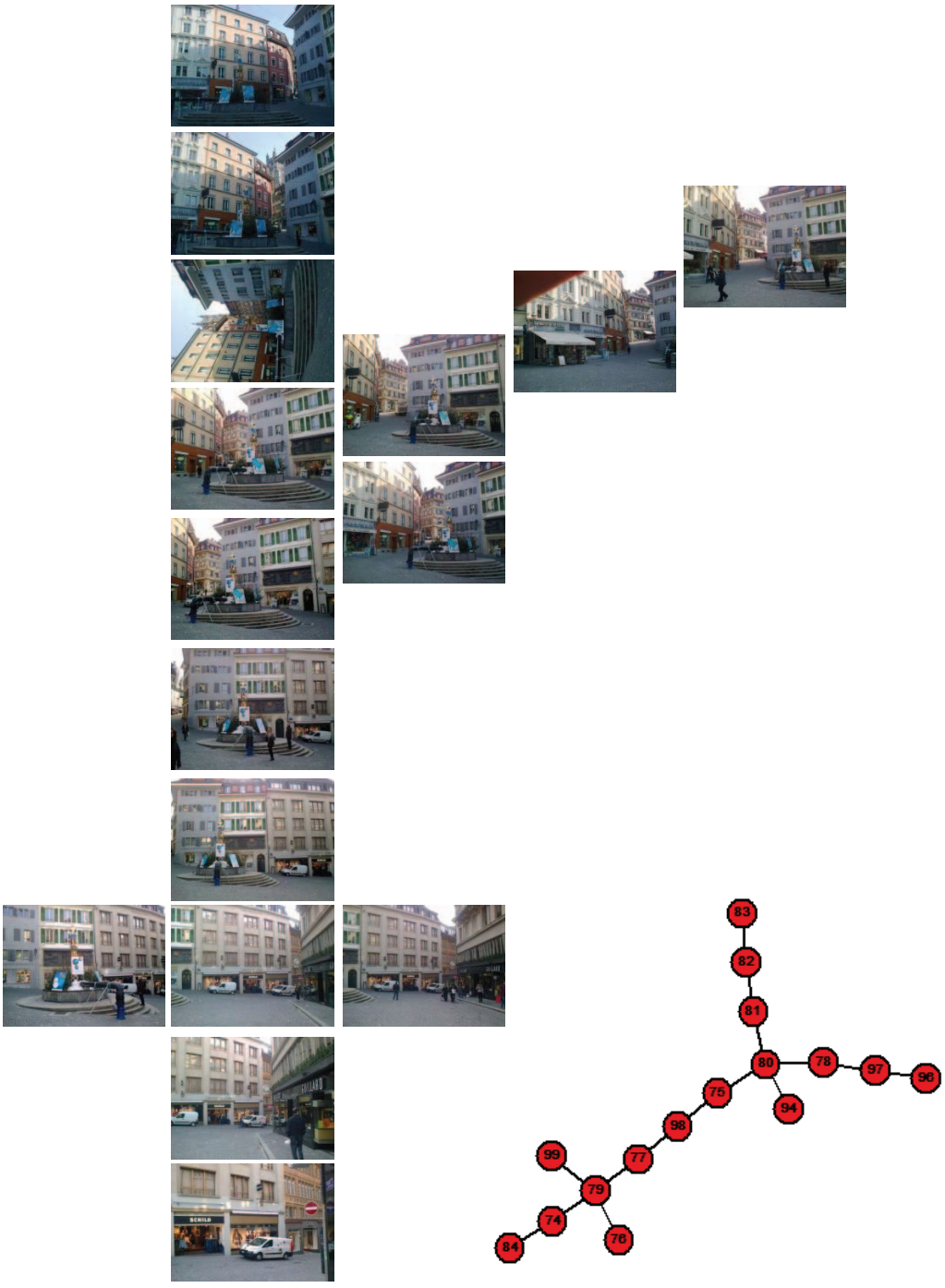


Figure 4.8: One of the basal trees for the Lausanne Dataset. This is one of densest collection of images.

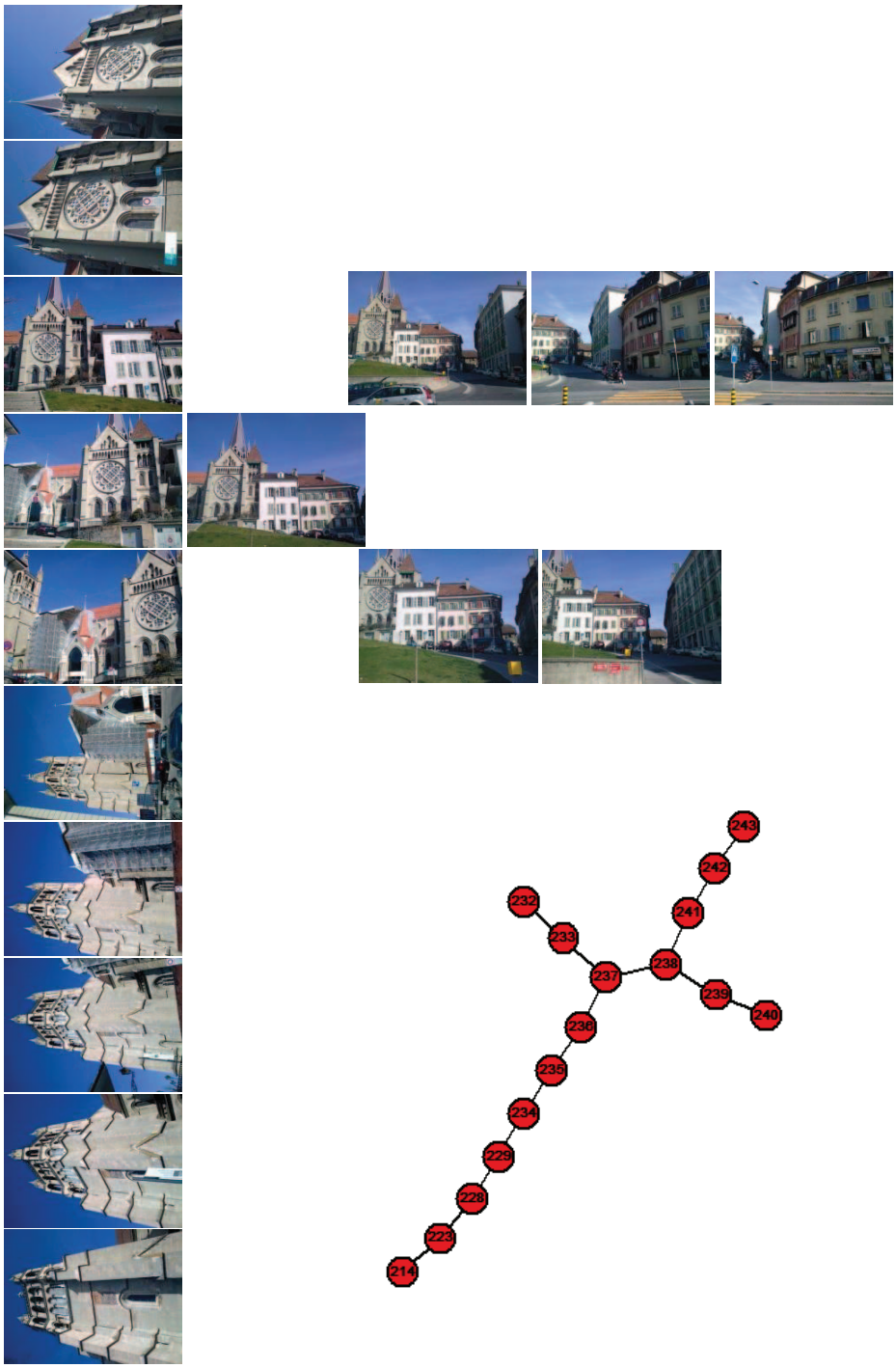


Figure 4.9: One of the basal trees for the Lausanne Dataset. Notice the significant difference between the leaf nodes in just few images.

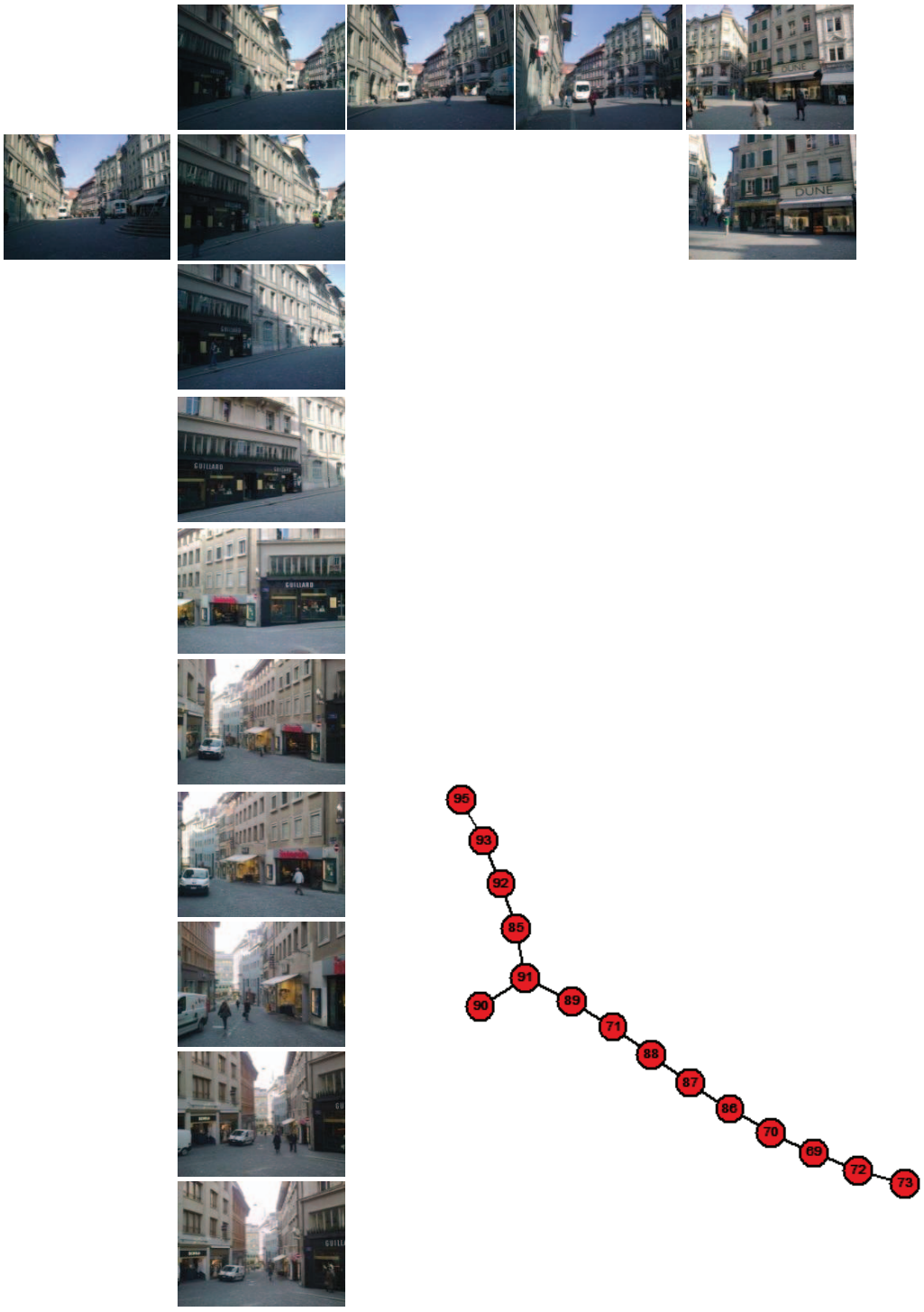


Figure 4.10: One of the basal trees for the Lausanne Dataset. Notice that this arrangement is almost linear and thus easy to follow. Shows how important linear arrangements can be.

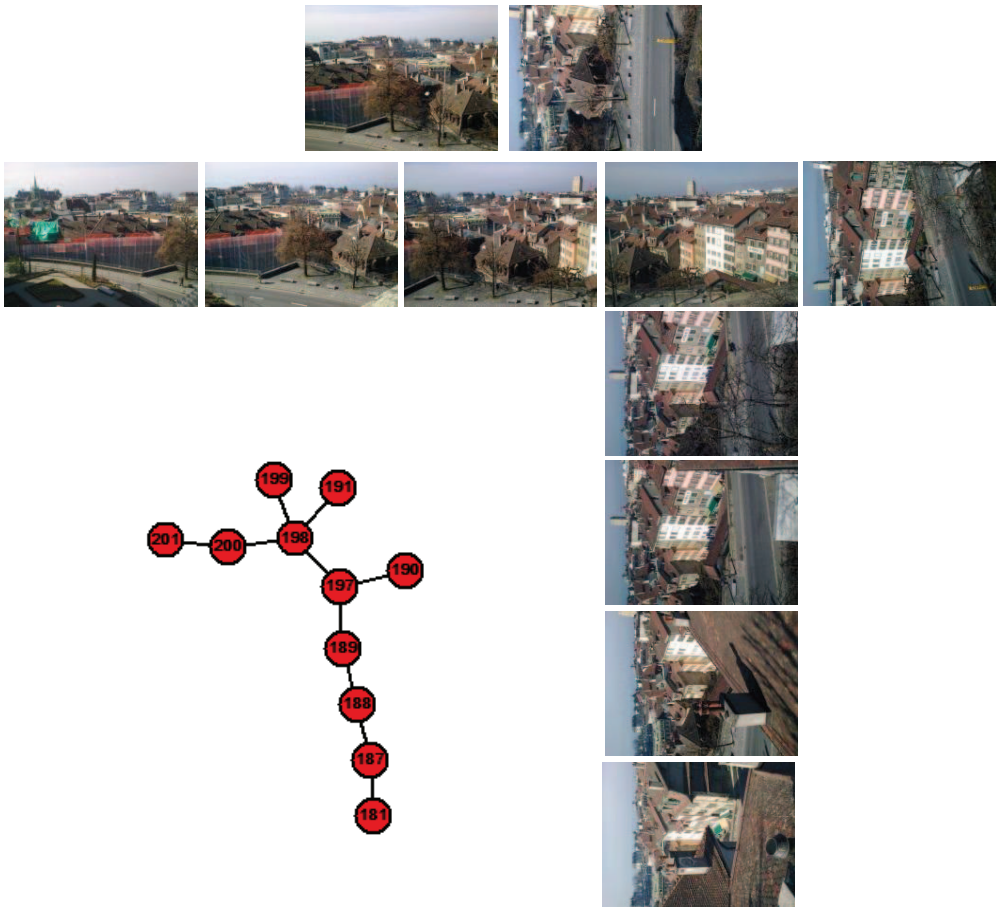


Figure 4.11: One of the basal trees for the Lausanne Dataset. Notice the significant changes between the leaves of the tree.

though sparingly. The images in the Lausanne dataset are separated by very-wide-baselines and are matching images in them are very difficult to identify using state-of-the-art methods. Demoset has fewer images that are visually connected but the visually connected images have very wide-baselines and are very difficult to match as well. Images in both these datasets are tagged with GPS sensor information and magnetometer sensor information. For the Lausanne dataset, images of calibration grid are provided to help in estimating the intrinsic calibration parameters. However, such data is not provided for the Demoset. We found Nokia dataset very interesting, particularly the Lausanne dataset which is well-connected but sparingly so much that even humans would find it difficult to see the

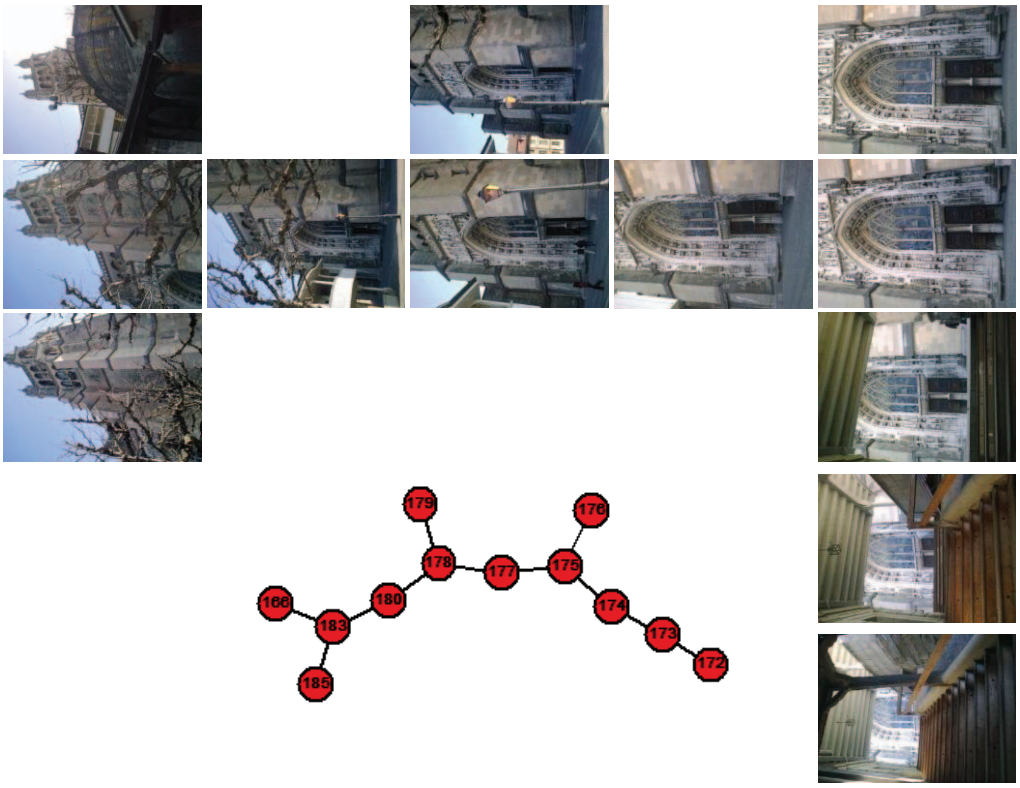


Figure 4.12: One of the basal trees for the Lausanne Dataset. Notice the significant changes between the leaves of the tree.

connections between the images. It poses an excellent challenge for research by pushing geometry based image organization problem to its limit.

The Oxford building [11] dataset is another challenging collection of 5063 images. These images have been collected from Flickr by searching various landmarks in Oxford using different queries. However, there are many images in the dataset which do not belong to any landmark. Also, there are plenty of images that have nothing to do with Oxford buildings and are not related to any other image in the dataset, making it a difficult and an interesting dataset. This dataset has been previously used for object retrieval, which can be seen as simpler version of problem of image organization. 5 wide-baseline images of 11 buildings were used as queries to evaluate performance of object retrieval system in the research that introduced this dataset. In our research, we do not use or assume any

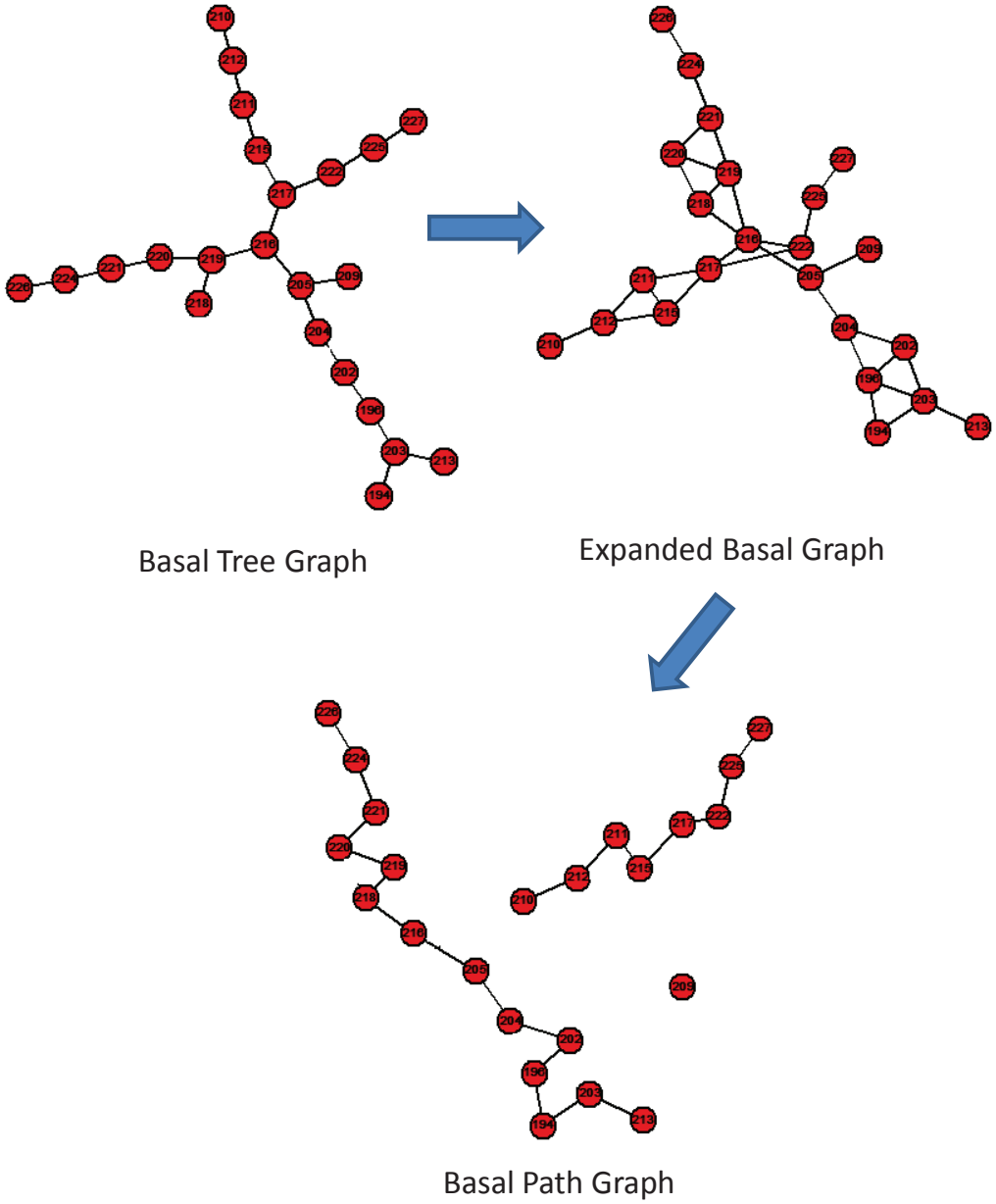


Figure 4.13: Expanded graph and basal paths in the basal tree shown in Figure 4.7. Note that the basal path graph split into 3 parts

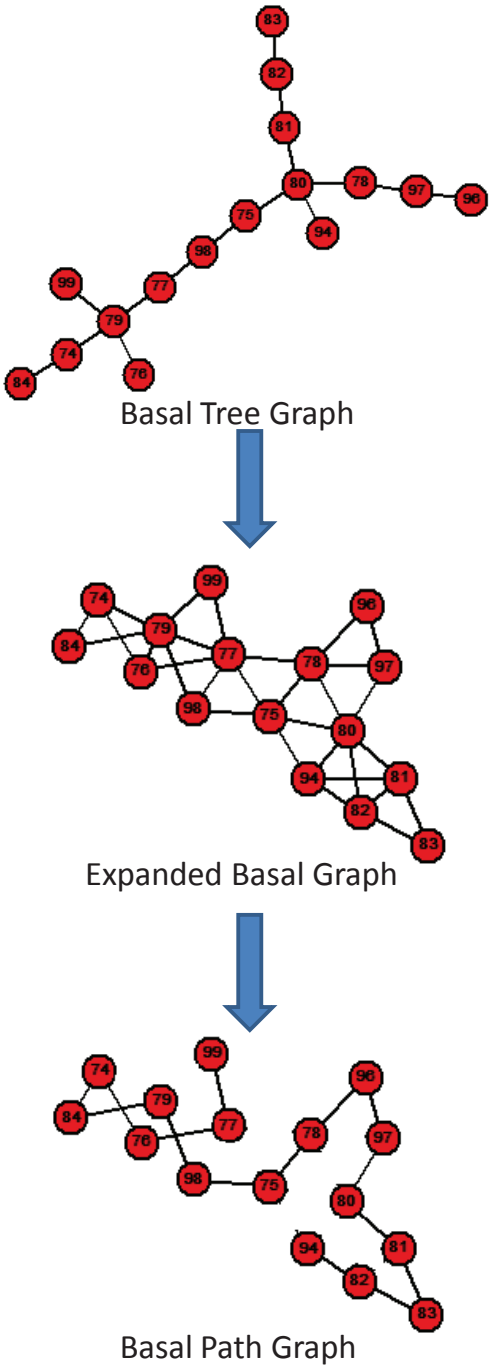


Figure 4.14: Expanded graph and basal paths in the basal tree shown in Figure 4.8. Note that the basal path graph in this case is fully connected.
labelpath3

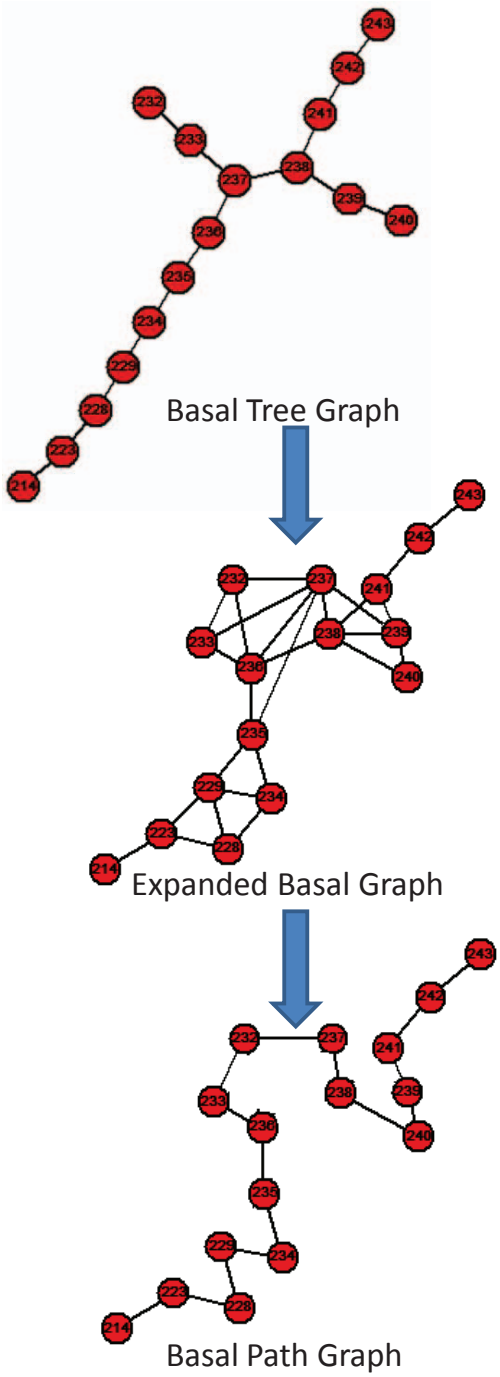


Figure 4.15: Expanded graph and basal paths in the basal tree shown in Figure 4.9.Note that the basal path graph in this case is fully connected.

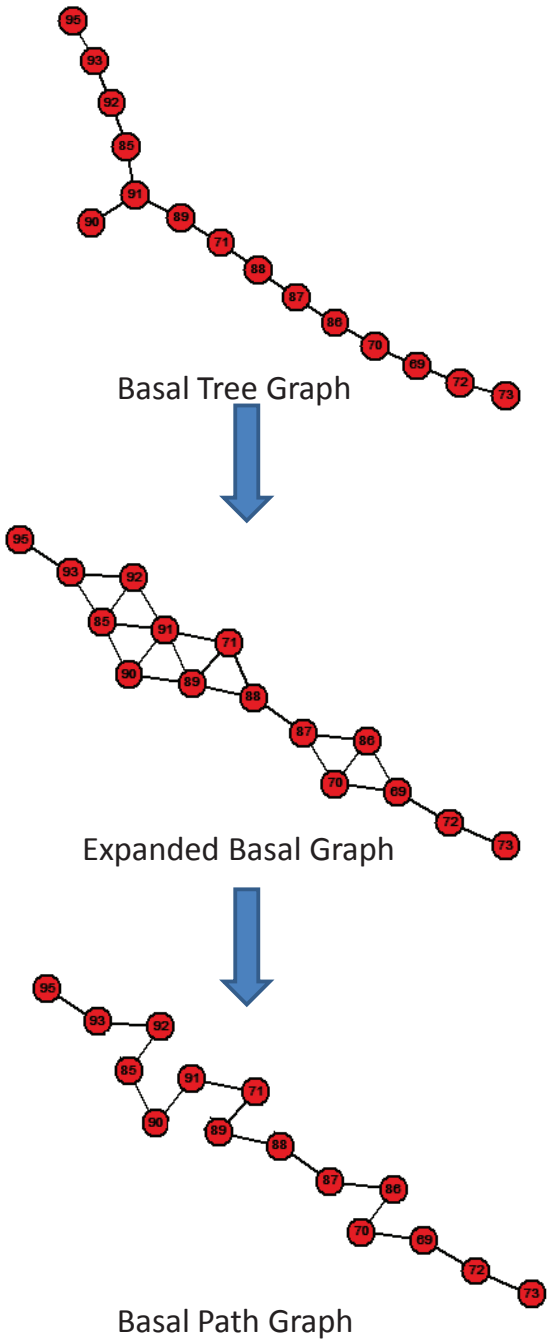


Figure 4.16: Expanded graph and basal paths in the basal tree shown in Figure 4.10. Note that the image that was outside got accommodated in the path

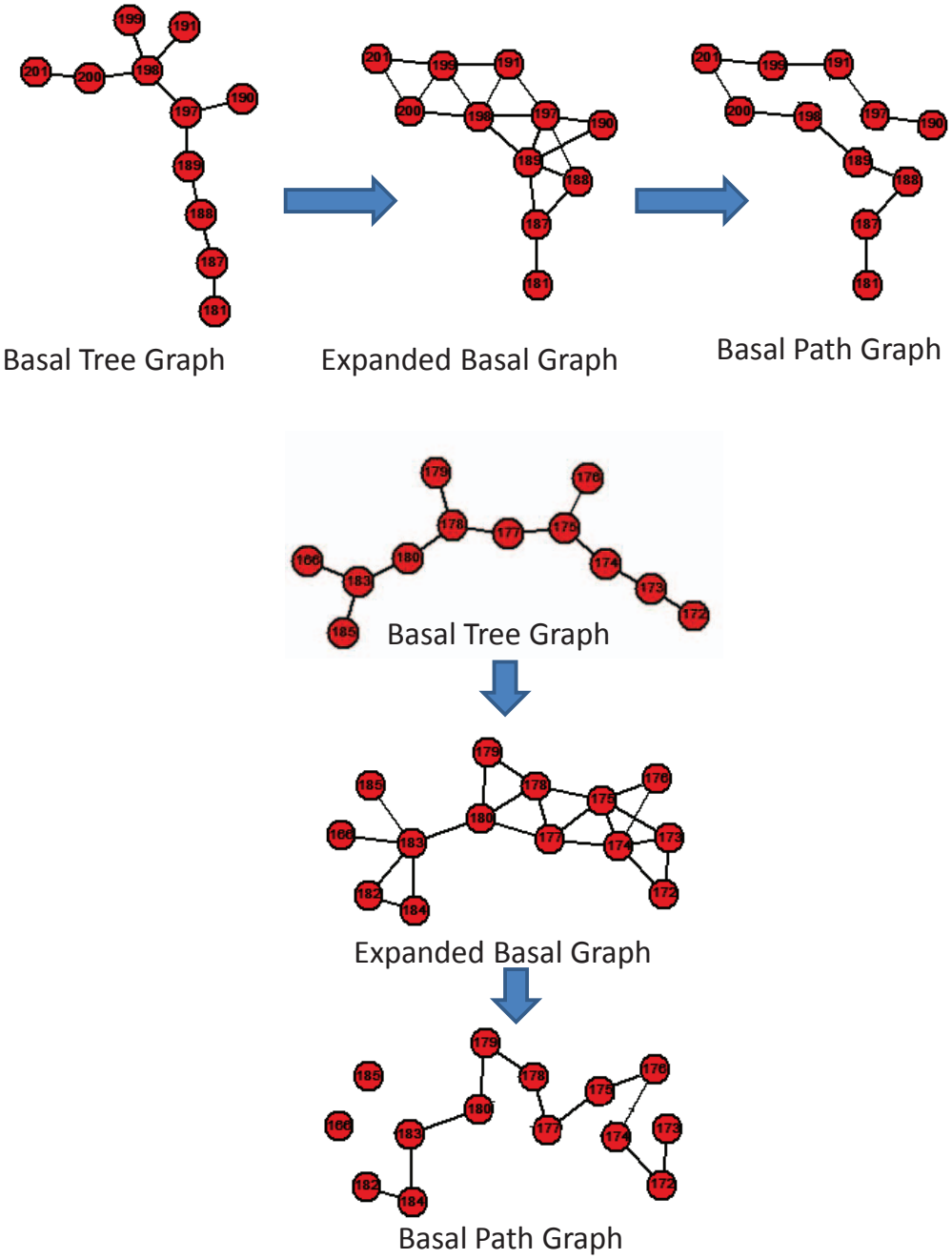


Figure 4.17: Expanded graph and basal paths in the basal tree shown in Figure 4.11 and Figure 4.12. Note that the path for Figure 4.12 has two single images that could not find a place in the larger path.

previous knowledge on the dataset. We perform wide-baseline geometry based matching in order to get accurate organization of images. In the Oxford dataset, there are many similar overlapping images as well as very wide-baseline images in the presence of images that do not have any match with others also known as distractors.

The ArtsQuad [42] dataset has 6514 images. All the images in this dataset have GPS information from a commercial phone (an iPhone 3G) and 348 of the images have tags from a highly accurate survey quality (10 cm error) differential GPS device. This dataset has images that are more overlapping in the scene content than the Nokia dataset and has been used for 3D reconstruction as well. We use the 348 images with highly accurate survey quality GPS data as ground truth for the GPS noise detection problem. This dataset is one of the large datasets we use for our research and it tests the scalability of our algorithm.

4.3.2 Ground Truth and Performance Evaluation

All pairs of images in a collection of 348 images by combining the Lausanne dataset and the Demoset were visually inspected for a match and a 0/1 matrix of size 348×348 is prepared as ground truth to test the basal tree graphs produced by our algorithm CODIMSEG.

Basal tree graphs are produced by various versions of our algorithm for evaluation. Note that basal tree graphs are produced by thresholding a single spanning tree after the CODIMSEG algorithm converges. Next the accepted edges in the spanning tree forming the basal tree graphs and the edges rejected by our algorithm are tested against the ground truth to verify if the acceptance and rejections were valid. The algorithm that produces most valid results is considered the best.

We compared various versions of our algorithm, CODIMSEG-CCS-50, CODIMSEG-CCS-100, CODIMSEG-GIST-50, CODIMSEG-GIST-100 and CODIMSEG-GIST-384 with the pure geometric algorithm. 50, 100 and 384 are the size of the smaller side of the image in pixels. 11 inlier rate thresholds were used ranging from 0 to 0.5 with increments of

0.05 and 9 thresholds on the number of inliers were used from 0 to 80 with increments of 10. Thus, there are 99 combinations of thresholds giving 99 points using which ROCs are plotted. The ROCs are given by the upper convex hull of the 99 points. The formulas used for True Positive Rate (TPR) and False Positive Rate (FPR) are given by Equation 4.3 and Equation 4.4 respectively where TP, TN, FP, FN are true positive, true negative, false positive and false negative respectively.

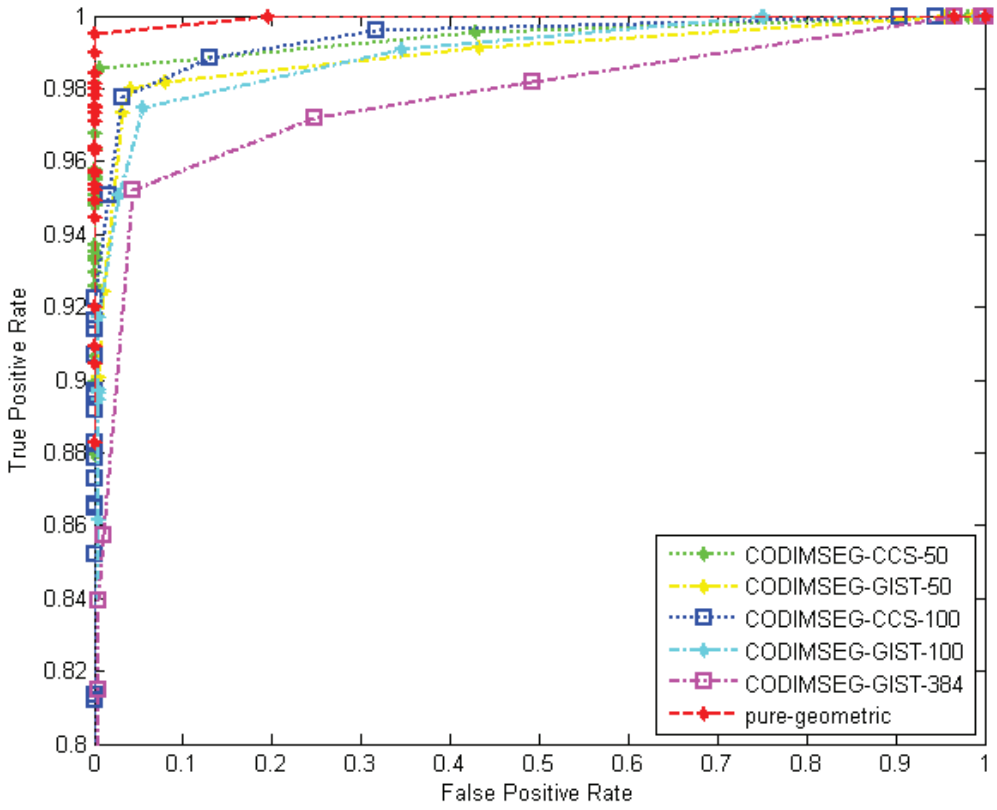


Figure 4.18: ROCs generated to verify the correctness of the geometry based thresholding to produce the basal tree graphs by different versions of our algorithm and the pure geometric version.

$$TPR = \frac{TP}{TP + FN} \tag{4.3}$$

$$FPR = \frac{FP}{FP + TN} \tag{4.4}$$

4.3.3 Results

The pure geometric algorithm performed the best followed by CODIMSEG-CCS-50, CODIMSEG-CCS-100, CODIMSEG-GIST-50, and CODIMSEG-GIST-100, CODIMSEG-GIST-384. 50 and 100 denote the size of the smaller side of the images in pixels. Surprisingly, the algorithm performs better with smaller images than with larger images. The performance of pure geometric is comparable with best version of our algorithm.

Next, we want to compare how successful the algorithms are in identifying connected images. Now, we do not consider True Negatives in the success of our algorithms. We now compare CODIMSEG with k-NN and k=3 and k=5 algorithm. CODIMSEG always converged in less than $3N$ geometric estimations for the dataset used for ground truth and thus k=3 or k=5 should be sufficient for this 348 image dataset.

We estimated the precision (PRE), accuracy (only considering true positive) (ACC), and 1- false positive rate for all the algorithms as shown in Table 4.1.

$$ACC = \frac{TP}{TP + FP + FN + TN} \tag{4.5}$$

$$PRE = \frac{TP}{TP + FP} \tag{4.6}$$

The results for all algorithms are found at the threshold at which pure geometric performs the best. In the Table 4.1, it can be seen that CODIMSEG-CCS-100 performs the best, even better than the pure geometric. This can be explained by the fact that geometric match scores are also just estimates like the photometric match scores, although they are considered the best. Hybrid algorithms provide photometric guidance and too which provides an additional advantage and thus has lower false positive rate. Note that pure

Table 4.1: Comparing various algorithms based on the ratio of true positives and total possible edges, and its product with precision and 1 - false positive rate.

Algorithm	Photometric Score	Image Size (Smaller side pixels)	True-Positives/ Total (A)	Precision (B)	1-FPR (C)	Product (AxBxC)
Pure Geometric	N/A	384	**0.7320**	0.9621	0.8925	**0.6285**
CODIMSEG	CCS	50	0.5965	0.9952	0.9927	0.5893
CODIMSEG	CCS	100	**0.6945**	0.9757	0.9400	**0.6370**
CODIMSEG	GIST	50	0.6254	0.9775	0.9600	0.5868
CODIMSEG	GIST	100	0.5620	0.9606	0.9456	0.5104
CODIMSEG	GIST	384	0.5821	0.9573	0.9323	0.5196
3NN	CCS	50	0.5159	1.0000	1.0000	0.5159
3NN	CCS	100	0.5476	0.9948	0.9935	0.5411
5NN	CCS	50	**0.6715**	0.9790	0.9561	**0.6285**
5NN	CCS	100	**0.6859**	0.9754	0.9450	**0.6322**

geometric is best in terms of accuracy. CODIMSEG outperformed 3NN by a good margin. However, 5NN gives equivalent results, but performs much more geometric estimations than CODIMSEG. Thus, CODIMSEG-CCS-100 is the winner in the table.

Chapter 5 Noisy GPS and Magnetometer Tag Detection

Images tagged with meta-data are often used for multi-view analysis. Any application using these meta-data would depend on their reliability. GPS and magnetometer pose are few among such meta-data, but their reliability can be under question. Thus, an analysis on the reliability of GPS and magnetometer tags for multimedia applications is required. In this work, we propose a method for detection of noisy GPS and magnetometer tags in N wide-baseline views using vision.

Given a collection of images, each with its GPS and magnetometer tags that record its location (latitude, longitude, altitude) and camera orientations (roll, pitch and yaw), we have to find the tags that are likely to be wrong. We propose two novel algorithms for detection of noisy GPS and magnetometer tags in multiple wide-baseline camera views. We call these algorithms, geometric voting and geometric eigen-voting. Our algorithm does not require the dataset to have a single connected visual path between all images. The algorithm identifies reliable vision estimates of rotation and translation between cameras and also outputs a measure of confidence on the correctness of the associated GPS and magnetometer tags. The reliable estimates can be identified by exhaustive estimation of epipolar geometry between all pairs of images or by estimating the expanded basal graphs as explained in previous chapter.

We show results on the Nokia Grand Challenge 2010 Dataset and the ArtQuad dataset. The Nokia dataset has 243 images in Lausanne dataset and 105 images in the Demoset. The Lausanne dataset is particularly useful since it also has the intrinsic calibration parameters along with the magnetometer and GPS tag. However, the Lausanne dataset does

not have ground truth. We use soft ground truth like epipolar lines and manual ordering of the images to help us judge if our overall approach is correct. In the initial tests, we found that magnetometer has very little noise and the GPS is the one that has significant noise. Thus, we completely shift our attention to the GPS noise detection problem. Next, we use the ArtQuad dataset that has images both with commercial GPS data and DGPS data. Unfortunately, the ArtQuad data does not have altitude information and the orientation information. However, the DGPS data in the ArtQuad dataset acts as a ground truth for benchmarking the performance of our algorithms. We also use synthetic data and vary the number of noisy data to get an idea of the percentage error that our algorithms can handle.

We consider the general case in which the camera views might or might not be connected. We consider two images to be connected if there is an intervening sequence of images that overlap with the adjoining ones in terms of their scene content. It would be possible to propagate geometric estimates among the visually connected views. Thus, identifying visual connectivity in the dataset is a part of the problem. When defining this connectivity we would like to use pairs of view that allow for the most reliable estimate of the geometric parameters. Such views should have high overlap and should have sufficiently separated baseline. The latter is not an issue in our dataset as the views are sparse. Instead of pairs of images, one could define connectivity in terms of view triples to allow for trifocal tensor estimates; however, in wide baseline datasets one is less likely to find good triples than good pairs. So, we restrict ourselves to considering visual connectivity in terms of pairs of images.

5.1 Background

The problem of refining GPS and magnetometer tags figured in Nokia Grand Challenge [6] in 2009 and 2010. So far, no research work has directly addressed this challenge problem using the dataset supplied with the challenge. The Nokia Grand Challenge dataset is a hard dataset containing images captured using mobile phone that has a camera equipped

with GPS and magnetometer sensors. The visual connectivity in this dataset is very sparse across views, unlike most other landmark scene datasets used in state-of-the-art research [13, 19, 52, 56, 115] for scene reconstructions and other applications [12, 61, 67, 73, 87, 101, 105, 112, 133, 135]. We use the Nokia Grand Challenge dataset and the ArtQuad dataset [42] in our research and consider the problem of detection of noisy GPS and magnetometer tags. After initial experiments, we realize that magnetometers are far more reliable than GPS. Thus, it is the GPS that is particularly of higher interest. We addressed the problem of detection [25] of noisy GPS and magnetometer tags in a collection of wide-baseline images.

5.1.1 GPS

GPS stands for Global Positioning System. It is a constellation of 24 geo-stationary satellites. 3 of the satellite should be able to see a point on Earth to report the coordinates of the point. However, the position reported by the GPS can be incorrect by several meters. There are multiple factors that can degrade the GPS signal. Broadly there are four reasons [4] namely, time taken by the signal to travel, clock errors, position errors and intentional degradation. GPS signals are slowed down by the Earth's atmosphere and by tall buildings forcing the signals to take longer paths. The receiver clock is not as accurate as the atomic clock in a GPS satellite leading to mapping errors. The position errors are due to inaccurately known position of a satellite with respect to ground or with respect to other GPS satellites.

5.1.2 DGPS

In open fields, GPS errors are within 15 meters and the average errors are close to 5 meters. The errors are more when the line of sight from the GPS satellites is interrupted. Thus, a better measurement scheme is required. DGPS stands for Differential Global

Positioning System. DGPS uses a reference base station with known coordinates to rectify the GPS. Thus, DGPS signals have less error. However, even DGPS data can have errors. In the dissertation, we have used DGPS to serve as ground truth for detection of GPS noise.

5.1.3 Magnetometer and Accelerometer

Magnetometer is an instrument used to measure the strength and sometimes the direction of a magnetic field. Often it is used to measure the magnetic field of the Earth. Accelerometers are used as motion sensors. Together with magnetometer, they can be used to sense the orientation of the camera as done in some modern phones like the Nokia 6210 Navigator. Magnetometer can get affected by external magnetic fields produced by electrical equipments around. In the dissertation, whenever we refer to magnetometer, we mean the orientation sensor that is magnetometer and accelerometer collectively.

We first present an exhaustive version of our algorithm and then suggest a modification to achieve a faster version.

5.2 Our Approach: Geometric Voting and Geometric Eigen-Voting

In our approach, we first find out the epipolar geometry estimates that are reliable enough to detect wrong GPS and magnetometer tags. All pair-wise estimates of epipolar geometry thus found, are decomposed to rotation and translation, and then they are compared with pair-wise magnetometer based rotation and GPS based translation. Rotation matrices are converted to Euler's angles. The highest cosine similarity between vision based translation and rotation and GPS based translation and magnetometer based rotation for each image relative to other images is indicative of the correctness of the GPS and magnetometer tag respectively in one version of our algorithm in Algorithm 6. In another version, that is, in Algorithm 7, we used the eigen-vector corresponding to eigen-value of 1 in

row-normalized cosine similarity matrix to indicate correctness of GPS and magnetometer tags respectively.

Epipolar geometry estimation between pairs of images [24, 125] is the costliest step of our approach. state-of-the-art epipolar estimation methods, especially those for wide baselines, rely on random sampling [47], which is computationally expensive. Exhaustive pair-wise geometry estimations would take $O(N^2)$ such estimations, where N is the number of views. Potentially matching features between an image pair form a putative correspondence set. We treat the rate of inliers to an epipolar geometry or the fraction of matches that satisfy the epipolar constraint as a measure of geometric similarity between two images. In the faster version of our algorithm, we reduce the number of epipolar geometry estimations by approximating this inlier rate among the putative correspondences between the image pairs by using CCS. We use the expanded basal graphs from the CODIMSEG algorithm in the faster version. Thus, as we can formulate confidence measures in the vision estimates, we estimate a measure of reliability of GPS and magnetometer tags in each image in the dataset except for the ones in which vision might fail to give us any reliable estimate.

Apart from the ground truth GPS positions in the ArtQuad dataset, we verify the correctness of our algorithm using soft ground truths like manual ordering of images and epipolar lines. Through experiments we show that the GPS tags from modern mobile phones with cameras can be outrageously wrong with respect to vision calibration. We performed exhaustive pair-wise estimation of fundamental matrices from GPS positions and magnetometer orientations and found how good they fit with the putative correspondences between the image pair. The result can be seen in Figure 5.1b. The images are manually ordered, so the high values are expected to be along the diagonals of the figure. The result shows that the fundamental matrices found using GPS positions and magnetometer orientations are far from correct. On the other hand, we see that vision based fundamental matrices fit the putative correspondences much better yielding much higher inlier rates as seen in Figure 5.1a. This shows that vision has a role to play. The faster version of our

algorithms that use the expanded basal graphs are briefly described in Algorithm 6 and Algorithm 7.

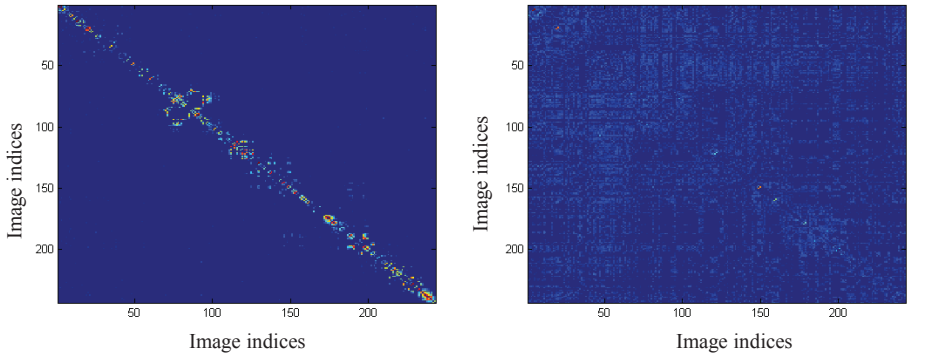
5.2.1 Problem Model, Notations and Mathematical Objective

Given N images $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_N$ with the ith camera having GPS position in geodetic coordinate system $\mathbf{P}_i^E = (l_i, g_i, h_i)$ ECEF (Earth Centered Earth Fixed) coordinates $\mathbf{P}_i = (x_i, y_i, z_i)$ and magnetometer based orientations $\mathbf{O}_i = (\psi_i^1, \psi_i^2, \phi_i^3)$, we need an estimate of confidence on the GPS positions and magnetometer orientations. This estimate of confidence can be later thresholded according to the requirement of particular applications, such as 3D view reconstruction or a photo tour through the views.

In order to get this confidence estimate using vision, we would estimate the degree of match between reliable vision based estimates of translations and rotations and corresponding GPS based translations and magnetometer based rotations between camera pairs. Our confidence estimate for an image is the maximum among all such estimates for an image paired with all other images. This is done separately for GPS and magnetometer sensors.

Algorithm 6 $\mathbf{C}^G, \mathbf{C}^M$ = GEOMETRIC-VOTING($\mathcal{V}$, Λ)

```
G(V, E = Λ)
for i = 1 → N − 1 do
  for j = i + 1 → N − 1 do
    if G_ij > 0 then
      Estimate unit vectors T^v_ij and R^v_ij using F_ij and K_ij
      Apply positive depth constraint to the rotation and translation pairs found.
      Estimate unit translation vectors and quaternions using T^g_ij and R^m_ij using GPS
      positions P_i = (x_i, y_i, z_i) and magnetometer orientations O_i = (ψ^1_i, ψ^2_i, ψ^3_i)
      D^G_ij = T^v_ij · T^g_ij
      D^M_ij = R^v_ij · R^m_ij
    end if
  end for
end for
C^G = max_j(D^G)
C^M = max_j(D^M)
```



(a) Rate of inliers to vision based F matrix in putative match between image pairs represented as entries of matrix converted to pixel values.

(b) Rate of inliers to GPS and magnetometer based F matrix in putative match converted to pixel values.

(c) Similarity between vision and GPS based translation represented as entries of matrix converted to pixel values.

(d) Similarity between vision and magnetometer based rotation represented as entries of matrix converted to pixel values.

Figure 5.1: (a) shows a matrix whose each entry represents the confidence of the vision based fundamental matrix, (b) shows the confidence of the fundamental matrix generated from the GPS and the magnetometer estimates, (c) shows the vision based confidence of the translation estimate from the GPS, (d) shows the vision based confidence of the rotation estimates from the magnetometer. The confidence in (a) and (b) are quantified by the inlier rates in the set of putative correspondences between image pairs. The images in the dataset are manually ordered in an approximate visually connected sequence. Thus, the high values are expected to be arranged along the diagonals. High values are towards red, low values are towards blue. This figure is best viewed in color.

Algorithm 7 φ_{ij} = GEOMETRIC-EIGEN-VOTING($\mathcal{V}$, Λ)

$G(\mathcal{V}, \mathcal{E} = \Lambda)$
for $i = 1 \rightarrow N - 1$ **do**
  for $j = i + 1 \rightarrow N - 1$ **do**
    if $\mathcal{G}_{ij} > 0$ **then**
      Estimate unit vectors $\mathbf{T}^v_{ij}$ and $\mathbf{R}^v_{ij}$ using $\mathbf{F}_{ij}$ and $\mathbf{K}_{ij}$
      Apply positive depth constraint to the rotation and translation pairs found.
      Estimate unit translation vectors and quaternions using $\mathbf{T}^g_{ij}$ and $\mathbf{R}^m_{ij}$ using GPS positions $\mathbf{P}_i = (x_i, y_i, z_i)$ and magnetometer orientations $\mathbf{O}_i = (\psi^1_i, \psi^2_i, \psi^3_i)$
      $\mathbf{D}^G_{ij} = \mathbf{T}^v_{ij} \cdot \mathbf{T}^g_{ij}$
      $\mathbf{D}^M_{ij} = \mathbf{R}^v_{ij} \cdot \mathbf{R}^m_{ij}$
    end if
  end for
end for
$\mathbf{C}^G$ = eigenvector corresponding to eigen-value of 1 in $\mathbf{normalize_j}(\mathbf{D}^G)$
$\mathbf{C}^M$ = eigenvector corresponding to eigen-value of 1 in $\mathbf{normalize_j}(\mathbf{D}^M)$

5.2.2 Geodetic to ECEF Coordinates Conversion

Given geodetic coordinates with latitude, longitude and altitude (l_i, g_i, h_i), its coordinates (x_i, y_i, z_i) in the ECEF coordinate system is given by the following equations. $r_\ominus$ and $e_\ominus$ are the radius and eccentricity of earth respectively. v_i is the radius of curvature in prime vertical [111]. Latitude and longitude are in radians, and h_i, (x_i, y_i, z_i), $r_\ominus$ and v_i are in meters. Interested readers are referred to [140] for detailed discussion on the ECEF coordinates.

$$v_i = \frac{r_\ominus}{(1 - e^2_\ominus sin^2 l_i)^{1/2}} \tag{5.1}$$

$$x_i = (v_i + h_i)\cos(l_i)\cos(g_i) \tag{5.2}$$

$$y_i = (v_i + h_i)\cos(l_i)\sin(g_i) \tag{5.3}$$

$$z_i = (v_i(1 - e^2_\ominus) + h_i)\sin(l_i) \tag{5.4}$$

5.2.3 Reliable Vision Estimates

Let $\mathbf{T}_{ij}^{v}$ and $\mathbf{R}_{ij}^{v}$ be the vision-based estimate of translation unit vector and Euler's angles and $\mathbf{T}_{ij}^{g}$ and $\mathbf{R}_{ij}^{m}$ be the GPS based translation unit vector and magnetometer based rotation respectively for an image pair $\{\mathcal{I}_i, \mathcal{I}_j\}$. Next, let $\mu(\mathbf{F_{ij}}|\mathbf{X_{ij}})$ be a probabilistic measure of the number of supporting correspondences for the vision based rotation and translation estimate among all putative correspondences. This is defined more precisely later. Let the number of all putative correspondences be $|\mathbf{X}_{ij}|$ between image pair $\{\mathcal{I}_i, \mathcal{I}_j\}$. Γ_1, Γ_2 and Γ_3 are thresholds on estimated number of inliers $\mu(\mathbf{F}_{ij}|\mathbf{X}_{ij})$ and lower and upper threshold on inlier rate $\frac{\mu(\mathbf{F}_{ij}|\mathbf{X}_{ij})}{|\mathbf{X}_{ij}|}$ respectively. The measure of reliability of vision estimates is give by :

$$G_{ij} = \left(\Gamma_3 \geq \mu(\mathbf{F}_{ij}|\mathbf{X}_{ij}) \geq \max\left(\frac{\Gamma_1}{|\mathbf{X}_{ij}|}, \Gamma_2 \right) \right). \tag{5.5}$$

If the number of inliers is less or if the inlier rate is too low, the epipolar geometry obtained is less likely to be accurate. At the same time, if inlier rate is very high, the scene is likely to be planar. We want to avoid epipolar geometry estimates from a pair of camera looking at a planar scene, or not looking at the same scene at all. Thus, we threshold out image pairs that have inliers less than Γ_1=20 and inlier rate less than Γ_2=0.25 or inlier rate greater than Γ_3=0.75. These thresholds are used on geometric estimates of inliers and inlier rate. Similar, thresholds are used in [12]. From here on, only the thresholded reliable vision estimates are used. Optionally, expanded basal graphs can be used as mentioned in the Algorithm 6 and Algorithm 7. In the following subsections, we discuss estimation of $\mathbf{T}_{ij}^{v}$ and $\mathbf{R}_{ij}^{v}$.

5.2.4 Vision-based Rotation and Translation

In order to estimate vision based rotation and translation between camera pairs, we first need to estimate a fundamental matrix $\mathbf{F}_{ij}$ for each pair of images. This pose-estimation is done using either the 7-points algorithms or the 8-points algorithm [65]. In presence of outliers, like in our case, pose-estimation is done in several iterations of 7-points algorithms or the 8-points algorithm in a random sample consensus framework.

In this work, we use the BLOGS algorithm [24] for estimation of the fundamental matrix. BLOGS stands for Balanced LOcal and Global Search and it is able to tolerate high outlier rate in epipolar geometry estimation, making it suitable for use in applications handling wide-baseline image pairs. BLOGS performs simultaneous local and global searches for the best motion model by sampling from distributions of correspondence probabilities. Global searches are done using a photometry based probability distribution which remains constant throughout, and local searches are done using 'best so far' geometry based probability distribution in the form of Joint Feature Distribution proposed by Triggs [128]. These two kinds of searches complement each other in finding a better motion model faster in successive iterations. Moreover, BLOGS algorithm is also able to avoid degenerate configurations [127] by checking the mutual scatter in each pair of correspondences amongst the ones used to generate a motion model. The mutual scatter should be above a decided threshold in each pair of correspondences for the motion model to be non-degenerate.

After finding the fundamental matrices $\mathbf{F}_{ij}$ using BLOGS, intrinsic calibration matrix $\mathbf{K}_i$ and $\mathbf{K}_j$ are used to estimate essential matrix $\mathbf{E}_{ij}$ from $\mathbf{F}_{ij}$. If all the images are captured from the same camera, $\mathbf{K}_i$ is equal for all i.

$$\mathbf{E}_{ij} = \mathbf{K}_j^T \mathbf{F}_{ij} \mathbf{K}_i \tag{5.6}$$

Further, matrix $\mathbf{E}_{ij}$ is decomposed into $\mathbf{T'}_{ij}$ and rotation matrices. This decomposition gives translation that is either negative or positive and two possible rotation matrices

leading to four possible combinations. The one that leads to positive depth of the scene captured in camera views, is the correct estimate of translation and rotation. The reader is referred to [66] for this decomposition. $\mathbf{T'}_{ij}$ is converted into unit translation vector $\mathbf{T}^v_{ij}$. Next, the rotation matrix is converted to quaternions $\mathbf{R}^v_{ij}$. For this conversion, the reader is referred to [113].

5.2.5 Tag Confidence Estimation

The dot product of corresponding unit translation vectors from GPS and vision give a confidence measure of the GPS tag involved. The product of cosine of the difference between Euler angles from magnetometer and vision gives a confidence measure of the magnetometer tag involved. Alternately, dot product between unit quaternions can be used to give confidence measure of the magnetometer tag.

$$\mathbf{D}^G_{ij} = \mathbf{T}^g_{ij} \cdot \mathbf{T}^v_{ij} \tag{5.7}$$

$$\mathbf{D}^M_{ij} = \mathbf{R}^m_{ij} \cdot \mathbf{R}^v_{ij} \tag{5.8}$$

$\mathbf{R}^g_{ij}$ and $\mathbf{R}^m_{ij}$ are the confidence measure of the GPS and magnetometer tags respectively. The maximum in the rows of the matrix $\mathbf{R}^g_{ij}$ and $\mathbf{R}^m_{ij}$ give us the respective confidence measure of GPS and magnetometer tags in our approach.

$$\mathbf{C}^G_i = \max_j \mathbf{D}^G_{ij} \tag{5.9}$$

$$\mathbf{C}^M_i = \max_j \mathbf{D}^G_{ij} \tag{5.10}$$

The Equation 5.9 and 5.10 are for GEOMETRIC-VOTING. Similarly, these equations can be modified for GEOEMTRIC-EIGEN-VOTING as shown in Algorithm 6.

5.2.6 GPS and Magnetometer based Fundamental Matrix

First, we calculate GPS based translations $\mathbf{T}_{ij}^{g}$ after conversion of GPS coordinates to ECEF coordinates. Next, we estimate the rotation matrix $\mathbf{R}_{ij}^{m}$ be $\mathbf{R}(\psi_{ij}^{m_z})\cdot\mathbf{R}(\psi_{ij}^{m_y})\cdot\mathbf{R}(\psi_{ij}^{m_x})$ where $\mathbf{R}(\psi_{ij}^{m_z}), \mathbf{R}(\psi_{ij}^{m_y}), \mathbf{R}(\psi_{ij}^{m_x})$ are the rotation matrix formed by Euler's angle in the order roll, yaw, pitch respectively. Next, magnetometer based rotations $\mathbf{R}_{ij}^{m}$ and GPS based translation $\mathbf{T}_{ij}^{g}$ are converted to a 3×4 projection matrix $\mathbf{PR'}_{ij}$. Another matrix $\mathbf{PR}_{ij}$ is formed as shown in following equations.

$$\mathbf{PR}_{ij} = \mathbf{K}_i[\mathbf{I}(3)|0] \tag{5.11}$$

I(3) stands for a 3×3 identity matrix. The 4th column of the matrix contains 0 entries as shown in the Equation 5.11.

$$\mathbf{PR'}_{ij} = \mathbf{K}_j[\mathbf{R}(\psi_{ij}^{m_z})\mathbf{R}(\psi_{ij}^{m_y})\mathbf{R}(\psi_{ij}^{m_x})|\mathbf{T}_{ij}^{g}] \tag{5.12}$$

For conversion of two projection matrices to a fundamental matrix, the reader is again referred to [66] [76]. We use this fundamental matrix to compare epipolar lines from vision based estimates with GPS and magnetometer sensor based estimates and to show result in Figure 5.1b.

5.3 Experiments

Accuracy and precision of our algorithms are measured on both real and synthetic data. Details of the data are in the following sub-section. Real data is the data that is coming from DGPS and GPS and images. Synthetic data are a set of point locations and corresponding noisy point locations.

Figure 5.2: Examples of the quality of epipolar line estimates based on GPS/magnetometer data and vision data for some image pairs from the Nokia dataset. The left column shows few reference points on one of the images pairs. The middle column shows the corresponding epipolar lines on the other image pair based on the GPS/magnetometer estimates. We do not see any epipolar line in one of the images above because the GPS and magnetometer estimates are so wrong that the corresponding epipolar lines are outside the image boundary. In the right column we see the epipolar lines based on the vision estimates.

In the Table 5.1, DGPS Vs GPS means that the translation vectors used to correct GPS has been estimated using the DGPS measurements for the sake of experiment. Vision Vs GPS indicates that vision based translation vectors were used as discussed in the chapter.

Synthetic data are a simulation of the DGPS and GPS data, assuming that the DGPS data is completely noise free. Random uniform noise is added to random points and the dataset is prepared. More details follow.

5.3.1 Datasets

Lausanne dataset in the Nokia dataset [6] is of 243 images with position and orientation tags as well as intrinsic camera calibration parameters. This allows it to be a good dataset for application of vision to detect noisy position and orientation tags. Unfortunately, the Lausanne dataset does not have any ground truth position and orientation. However, the overall disagreement of vision with position and orientation tags can be found using the Lausanne dataset. After initial experiments, it was realized that orientation tags are mostly correct and it is the position (GPS) data that needs to be studied.

ArtQuad dataset is another dataset we use that has the survey quality (10 cm error) DGPS data along with GPS data in 348 of its images. This dataset however does not have altitude information in the position tag and also does not have the orientation tag.

Other than these two datasets, we used synthetic datasets of 1000 2D points with coordinates between (0,0) and (100,100) and perturbed different %ages of data point varying from 10% to 70% with uniformly random noise ranging up to (10,10). In the synthetic data, no noise was added to the reference data. In another experiment, DGPS based translation vectors were also used as a replacement for vision based translation vectors. In the last experiment, vision based translation vector was used.

5.3.2 Ground Truth and Performance Evaluation

Ground truth noise is found between GPS and DGPS. Any tag with noise above 4m is considered a noisy tag for the real data. For the synthetic data, we have to identify all the data to which noise has been added synthetically. Accuracy and precision values for the detection of noisy tags were found for the synthetic and real dataset.

The confidence measures returned by the algorithm are thresholded to find the noisy tags. In Algorithm 6, the threshold is done below mean, and in Algorithm 7, the threshold is kept below standard deviation minus mean of the confidence. Actual noise is found using the difference between GPS and DGPS. The detections are compared and the accuracy and precision are found.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.13}$$

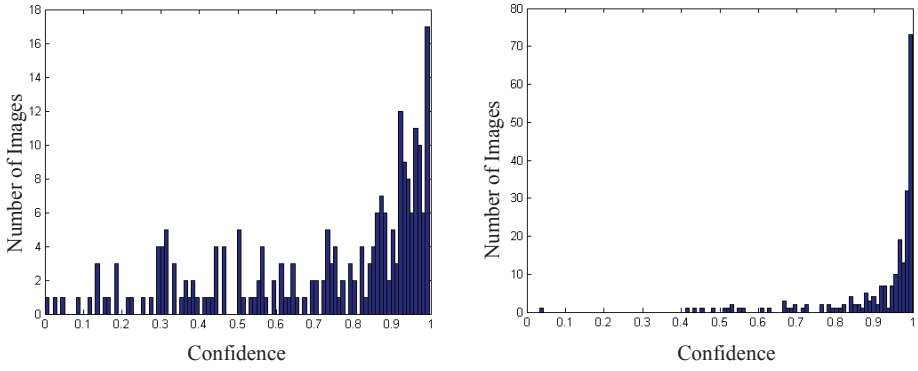
$$PRE = \frac{TP}{TP + FP} \tag{5.14}$$

5.3.3 Results

All 243 images from the Lausanne dataset were sub-sampled to size of 50×67 for evaluation of photometric similarity matrix. Pose-estimates were evaluated at size of 384×512 for these images.

We can only detect noisy GPS and magnetometer tags, where our vision estimates are good. Since the problem is new, our concern is accuracy primarily and then speed. So, we also performed exhaustive pairwise analysis apart from just our fast algorithm. Moreover, this also helped us in analysis of speed gain and accuracy loss of our fast algorithm over the exhaustive algorithm.

Further, in Figure 5.1c, we show the cosine similarity between exhaustive pairwise GPS based translation unit vectors and vision based translation unit vectors, that we estimate by finding a dot product between the two. It can be seen in the figure that higher values are more towards the diagonal as expected because the images in the dataset are manually arranged in an approximate order. It can also be noticed that high values in Figure 5.1c are spread in patterns across the similarity matrix. This is apparently because translation derived from a low quality (with regards to number of inliers to it) fundamental matrix (which we threshold out) has a moderate (neither good nor bad) match with GPS based translation. Next, in Figure 5.1d, we show the product of the cosine of the absolute values of difference between Euler's angles from the magnetometer and vision for all pairs of images. In the figure, we notice that high values are very close to the diagonal and rest of the values appear as random Guassian noise. Not only that, the high values are close to the maximum possible. This is closer to what ideally one would expect. This shows that magnetometer readings are quite accurate, and strengthens the qualification of vision estimates as a judge to decide the noisy GPS tags.

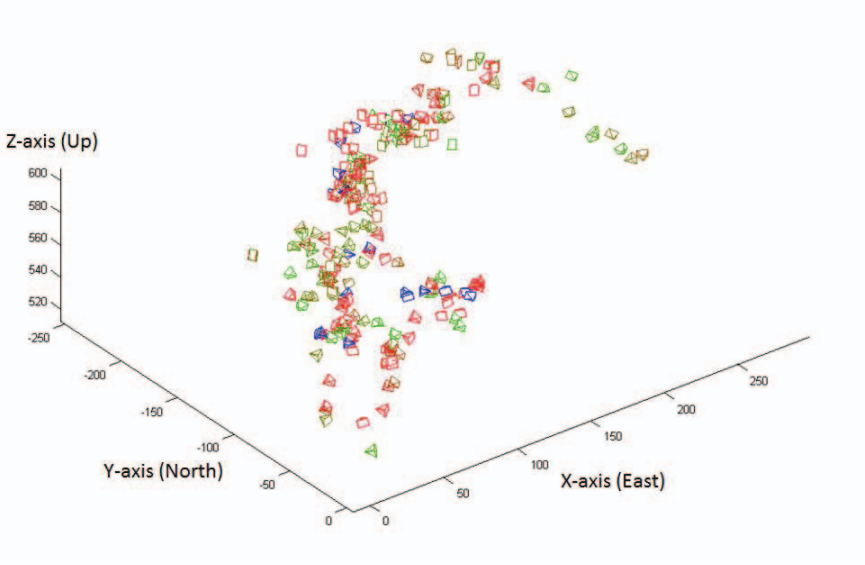


(a) Histogram of GPS confidence

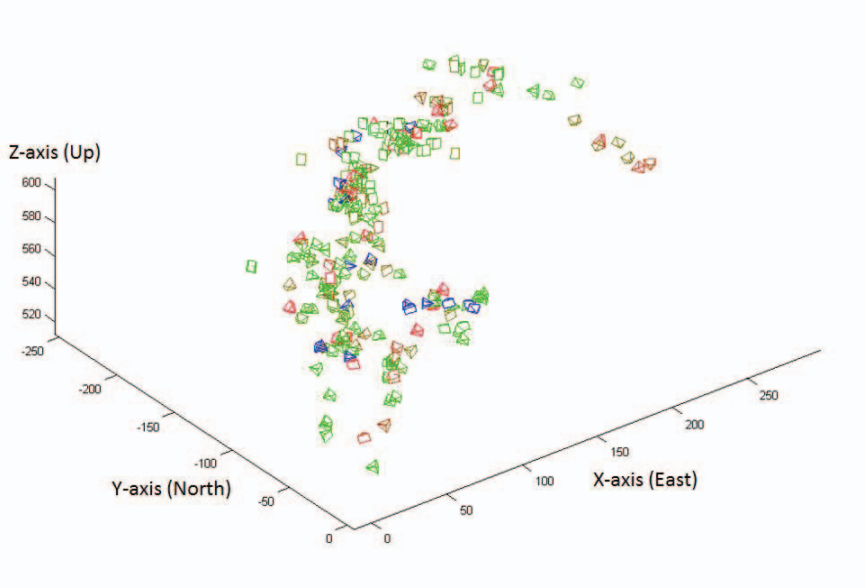
(b) Histogram of Magnetometer confidence

Figure 5.3: Histogram of vision based confidence on magnetometer and GPS based pose estimates.

In Figure 5.2, we show the difference between estimates of fundamental matrix from vision and that from magnetometer and GPS. The epipolar lines in the corresponding image



(a) GPS error detections



(b) Magnetometer error detections

Figure 5.4: Vision based detection of noisy magnetometer and GPS based epipolar geometry estimates. Figure shows GPS locations of cameras in local ENU(East North Up) coordinates with negative Y-axis as north and positive X-axis as east. The reference camera is at the origin. In (a), we show the noise detections in the GPS estimates of the cameras using colors. In (b), we show the noise detections in magnetometer estimates of the cameras using colors. In both (a) and (b), greener cameras are close to correct, while more red cameras are more noisy. Blue colored cameras denote the ones for which vision cannot be used to detect reliability of the GPS or magnetometer tags. All other colors like yellow and orange are formed by combination of red and green.

should pass exactly through the point corresponding to marked points on the first images. While the vision based estimates were found to be accurate, the GPS and magnetometer based fundamental matrices were found to be far from accurate, except for the first image pair. The first image pair is the one that has highest support for its GPS and magnetometer based fundamental matrix in the entire dataset. This experiment together with the Figure 5.1b clearly brings out the problem and the role vision can play in this regard.

In Figure 5.3a, we show a histogram of our estimated GPS tag confidence measure given in Equation 5.9. Ideally all the values should have been close to 1. This shows that there are many noisy GPS tags. In Figure 5.3b, we show a histogram of our estimated magnetometer tag confidence measure given in Equation 5.10. Most of the values are close to 1. This shows that most of the magnetometer tags are nearly accurate. In both Figure 5.3a and Figure 5.3b, we can see that the distribution of low values in the histogram is Gaussian like, which fits our expectation since for the tags which are incorrect we might have any degree of match between the vision and GPS or magnetometer sensor measurements. Similarly, high values are centered near 1.

In Figure 5.4, we show our results in color codes on GPS and magnetometer position and orientation respectively shown in East-North-Up coordinate system with the reference image at the origin. C_i^g and C_i^m are used to color code the confidence measures for display. These values range between -1 and 1. -1 means that vision cannot detect whether the GPS or magnetometer estimate is noisy because there was no reliable vision estimate at all to determine that and such cameras are shown in blue. Other values below 0 are shown in red. Values between 0 and 1 are shown in colors ranging from red to green. Green is closer to 1.

$$\mathbf{CC}_i^G = 1 - 4\frac{\cos^{-1}(\mathbf{C}_i^G)}{\pi} \tag{5.15}$$

$$\mathbf{CC}_i^M = 1 - 4\frac{\cos^{-1}(\mathbf{C}_i^M)}{\pi} \tag{5.16}$$

Cameras with less noise in their tags are greener and the ones with higher noise are redder. The images for which, vision cannot decide whether the tags are correct or not are colored in blue.

Table 5.1: Accuracy and precision values indicating how the proposed algorithms performed in identifying noisy GPS tags. The experiments were performed on synthetic data by varying the percentage of noisy GPS from 10% to 70%. On real data from ArtQuad dataset, ground truth DGPS was used to estimates the unit transaltion vectors to judge the GPS data, and in another experiment, vision based unit translation vectors were used.

		Geometric Voting		Geometric Eigen Voting	
Type of data	% Error Value	Accuracy	Precision	Accuracy	Precision
Synthetic Data	10%	97.18%	100.00%	96.40%	89.41%
	20%	91.62%	100.00%	90.07%	94.21%
	30%	85.26%	100.00%	82.16%	95.51%
	40%	78.20%	100.00%	71.62%	91.03%
	50%	71.25%	100.00%	60.79%	85.99%
	60%	63.42%	100.00%	49.57%	80.00%
	70%	54.43%	100.00%	37.28%	73.24%
Real Data	DGPS Vs GPS	65.42%	100.00%	55.33%	21.05%
	Vision Vs GPS	63.83%	76.66%	68.30%	74.47%

From the Table 5.1, we can see that the precision of the algorithm 6 is remarkable. It did not perform well only when real vision data was used, understandably because the images are sparsely connected. Performance of DGPS Vs GPS and comparing it the results from the synthetic data, one might conclude that the GPS data has errors about 60% of the time. Eigen-voting always performs worse that the simple geometric voting algorithm that uses a simple strategy of accepting the maximum of all votes.

Chapter 6 Discussion and Conclusion

In this dissertation, we addressed the problem of leveraging geometric information in wide-area sparse-view datasets by geometry based image organization in the form of basal graph structures. Such datasets have been used less in research. We used the Nokia, Oxford Building, ArtQuad dataset and twenty other wide-baseline image pairs. Nokia dataset is a very challenging dataset and has never been used except by us. To test our algorithms, ground truth was done on the Lausanne dataset by visually looking at the matches. Correct correspondences in the 20 image pairs were manually identified and hand-marked points on pairs of images were used to test result from the image matching and feature matching algorithms. ArtQuad Dataset has GPS ground truth that was used to test the algorithm for detection of noisy GPS and magnetometer tags.

Towards solving the problem we have five main contributions. First, we proposed a hop-diffusion search mechanism for a local and global search. We used this approach to search for a non-degenerate estimate of epipolar geometry. In terms of impact on the epipolar geometry problem, we find that this approach results in significantly better performance on images with wide baselines, scale changes, and repetitive structure. The success of the BLOGS approach can be attributed to four aspects: the photometry based global search, the JFD based geometric local search, the degeneracy criterion to weed out degenerate correspondences, and the M-estimator form of estimate of quality of a fundamental matrix. The influence of the mixing parameter α and degeneracy parameter β in BLOGS was studied and it was found that a more equal mixing is better and a low value of β results in better performance. BLOGS was compared with NAPSAC,

MAPSAC and BEEM. BLOGS uses 10 times lesser number of iterations than the best of the compared algorithms. The performance of BLOGS after 500 iterations is not significantly different from the best of the other algorithms in 5000 iterations. BLOGS has a linear time complexity over the number of iterations. NAPSAC and MAPSAC have worse than linear complexity. The time per iteration of BLOGS did not increase with the number of iterations but for NAPSAC and MAPSAC it increased. The cost of each iteration of BLOGS is lesser than the other compared algorithms. BEEM has the highest cost per iteration. BLOGS initially has a higher cost per iteration than NAPSAC and MAPSAC for number of iterations less than 500 due to the pre-processing time needed to calculate mutual scatter for estimating degeneracy, then later the time per iteration of BLOGS remains constant and for NAPSAC and MAPSAC it increases steadily. BLOGS identifies higher percentage of inliers within a pixel distance threshold than the compared algorithm in high number of iterations. In a dataset of 20 image pairs, the median Sampson's distance was found for various threshold of 1, 2 , 4 , 8 pixels for all algorithm executed for 50000 iterations. BLOGS identified the highest percentage of inliers. BEEM was the most inconsistent of all the compared algorithms. Performance of MAPSAC was the best amongst the rest of the compared algorithms.

Second, we propose a geometric match score from BLOGS and photometric match score CCS (Cumulative Correspondence Score) between image pairs using SIFT [77] features, although other features can also be used. The proposed photometric scores are a fast approximation of the computationally expensive geometric scores. These scores are an estimate of the inlier rate and number of inliers in putative correspondence sets between pairs of images. This is based on the observation that the outlier rate in a putative correspondence set increase with increase in geometric transformation. CCS was trained on the matching images in 'Demoset' of the Lausanne dataset.

Third, we use the photometric scores and the geometric scores to find groups of related images and to organize them in the form of basal tree graph structures using a novel hybrid algorithm called CODIMSEG (COnnected component DIscovery by Minimally

Specifying an Expensive Graph). The objective of the algorithm is to minimize the number of geometric estimations and yield results similar to what would be achieved if all-pair geometric matching were done. We found that CCS is significantly better than GIST and CODIMSEG is significantly better than k-NN in the problem of image organization. Using CCS, CODIMSEG converges faster. Our hybrid method does not fix the neighborhood search like k-NN does for geometric verification. We found that CODIMSEG-CCS-100 is the best performing algorithm. We found that all the versions of the CODIMSEG algorithm performed close to the pure geometric in terms of identifying matching images.

Fourth, we proposed a strategy for graph expansion that does not do exhaustive geometric estimations of all transitive closures as done in the state-of-the-art but selects those that connect the graph best as a spanning tree. We use the expanded graphs for finding basal paths using Hamiltonian Path approximation algorithms and also for detecting noisy tags. Conversion of a tree structure to a path structure might lead to splits. We qualitatively saw that the graph expansion helps in decreasing the number of splits.

In our third and the fourth contribution, we were lead to the same strategy while researching for advanced methods. Graph expansion algorithm re-uses the CODIMSEG algorithm again for minimizing the number of geometric estimations. In the state-of-the-art, image organization is done by photometric clustering using k-NN, seeking spanning trees and then graphs expansion. In our algorithm, we proposed a method that unifies the process of clustering, connectivity and expansion.

We addressed a novel problem of detection of noisy GPS and magnetometer tags in multiple views that are widely spaced in terms of rotation and translation between them. We proposed two versions of our algorithm - one performing exhaustive pair-wise epipolar geometry estimates and other performing epipolar geometry estimates opportunistically, which is the costliest step in our algorithm. Experiments done in this work clearly bring out the problem and the role vision can play. For all reliable estimates of vision, we estimated their match with GPS and magnetometer and found the best match which gives a robust estimate of tag confidence. We did this using two methods, geometric voting

and geometric eigen-voting. Geometric voting does not take the dependencies of votes into account. Geometric eigen-voting takes the dependencies of votes into account. Experiments done using the algorithms on the Nokia Dataset and the ArtQuad Dataset show that there is lot of noise specially in GPS tags. Thus, they cannot be used as initialization in vision but this noise can be detected using vision since vision is much more precise and accurate. The noise in GPS tags is found to be far more than the noise in magnetometer tags. After initial research, we concluded that magnetometer information is mostly correct and concentrated only on the GPS. On synthetic and real data, we found that geometric voting performed remarkably well in identifying noisy GPS tags in terms of precision. Overall, geometric voting performed better. Apart from the actual ground truth, epipolar lines and approximate manual ordering of the dataset served to verify correctness of our approach and result. The next obvious question is whether vision can be used to rectify these GPS tags. This is our future interest. While magnetometer tags can be rectified as it is, GPS tag rectification using motion estimates without refinement over multiple views is hard since vision based methods estimate translation only up to a scale.

Overall, BLOGS stands tall among all competing algorithm. It makes a place in a very highly researched class of algorithm. CCS provides a scheme suited for geometric applications that is both fast and uses less memory as it uses very small images. CODIMSEG is the highlight of this dissertation as it was used for basal graph discovery and later reused for basal graph expansion. CODIMSEG very smartly and neatly unifies the objectives and approaches in a single strategy. Geometric-voting and geometric-eigen voting are our algorithms for a new problem that has never been solved. Geometric voting is the simpler version that performed remarkably in terms of its precision.

6.1 Future Works

Geometry based image organization is the center of many applications one of which is 3D reconstruction. However, doing a 3D reconstruction from views that might have

come from arbitrary cameras that do not have the focal length tag is a big challenge. We have explored the field of self-calibration also known as auto-calibration and found that the state-of-the-art is far from making this a practical possibility. If reconstruction could be done without knowing any camera parameters after organizing the images, one would not have to collect images with focal length tags for reconstruction as done in [13, 52].

Another possibility is to organize videos according to their content. If a video is compressed into a single image, one would be able to use the solution suggested for image organization in the dissertation in this problem. Similarly, the video retrieval can be seen as a problem of image retrieval.

Vision-guided navigation is another interesting future work for me. If we start at a point and keep taking connected pictures on our mobile phone, can we find out where we are exactly relative to where we started from. Interesting question is that would there be any scaling problems as we know that vision can only map the 3D world by a similarity transform. The scale is lost.

One other exciting work is solving the correspondence problem in order to identify galaxies or stellar configurations. In such cases, exhaustive point matching is impossible. The challenge is to search for a tiny set of dots in a big search space. There are plenty of other directions too that might get lit up in future.

List of References

[1] Concorde home downloads. http://www.tsp.gatech.edu/concorde/downloads/codes/cygwin/linkern.exe.gz.

[2] David Liebowitz's home page. http://www.robots.ox.ac.uk/~dl/home.html.

[3] gaimc : Graph algorithms in matlab code. http://www.mathworks.com/matlabcentral/fileexchange/5355-toolbox-graph.

[4] Garmin — What is GPS? http://www8.garmin.com/aboutGPS/.

[5] Ian Shimshoni's BEEM Code Page. http://mis.hevra.haifa.ac.il/~ishimshoni/BEEM/.

[6] Nokia Challenge 2010: Where was this Photo Taken, and How? http://comminfo.rutgers.edu/conferences/mmchallenge/2010/02/10/nokia-challenge/.

[7] PCA-SIFT. http://www.cs.cmu.edu/~yke/pcasift/mod_lowe_demoV2.tar.gz.

[8] Toolbox graph. http://www.mathworks.com/matlabcentral/fileexchange/5355-toolbox-graph.

[9] Visual Geometry Group Home Page. http://www.robots.ox.ac.uk/~vgg/data/data-aff.html.

[10] WBS Image Matcher. http://cmp.felk.cvut.cz/~wbsdemo/demo/.

[11] The Oxford Buildings Dataset. http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/, 2007.

[12] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: Exploring image collections in 3D. In *Proceedings of ACM SIGGRAPH*, pages 835–846. ACM, 2006.

[13] S. Agarwal, N. Snavely, I. Simon, S. Seitz, and R. Szeliski. Building rome in a day. In *Proceedings of ICCV*, pages 72–79. IEEE Computer Society, 2009.

[14] M. Aly. *Searching Large-Scale Image Collections*. California Institute of Technology, Pasadena, CA, USA, June 2011.

[15] M. Aly, M. E. Munich, and P. Perona. Indexing in large scale image collections: Scaling properties and benchmark. In *WACV*, pages 418–425. IEEE Computer Society, 2011.

[16] D. Applegate, W. Cook, and A. Rohe. Chained lin-kernighan for large traveling salesman problems. *INFORMS J. on Computing*, 15(1):82–92, Jan. 2003.

[17] X. Armangue and J. Salvi. Overall view regarding fundamental matrix estimation. *Image and Vision Computing*, 21(2):205–220, 2003.

[18] N. G. Arnaud Doucet, Nando de Freitas. *Sequential Monte Carlo methods in practice*. Springer, New York, NY, USA, 2001.

[19] Y. Avrithis, Y. Kalantidis, G. Tolias, and E. Spyrou. Retrieving landmark and non-landmark images from community photo collections. In *Proceedings of ACM Multimedia*, pages 153–162, Firenze, Italy, 2010. ACM.

[20] A. M. Baumberg. Reliable feature matching across widely separated views. In *Proceedings of CVPR*, pages I: 774–781, 2000.

[21] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, June 2008.

[22] J. L. Bentley. Fast algorithms for geometric traveling salesman problems. *INFORMS Journal on Computing*, 4(4):387–411, 1992.

[23] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B*, pages 259–302, 1986.

[24] A. Brahmachari and S. Sarkar. BLOGS: Balanced LOcal and Global Search for non-degenerate two view epipolar geometry. In *Proceedings of ICCV*, Kyoto, Japan, 2009. IEEE Computer Society.

[25] A. Brahmachari and S. Sarkar. Fast detection of noisy GPS and magnetometer tags in wide-baseline multi-views. In *Proceedings of ACM Multimedia*, pages 997–1000. ACM, 2011.

[26] A. Brahmachari and S. Sarkar. View Clustering of Wide-Baseline N-Views for Photo Tourism. In *Proceedings of SIBGRAPI*, Maceió, 2011. IEEE Computer Society.

[27] A. Brahmachari and S. Sarkar. Hop-diffusion monte carlo for very wide baseline epipolar geometry estimation. *IEEE Transaction on PAMI (accepted)*, 2012.

[28] A. S. Brahmachari. *BLOGS: Balanced Local and Global Search for Non-degenerate Epipolar Geometry, Master's thesis*. Dept. of Computer Science & Engineering, University of South Florida, Tampa, 2009.

[29] S. Brandt. On the probabilistic epipolar geometry. *Image and Vision Computing*, 26(3):405–414, March 2008.

[30] M. Brown and D. G. Lowe. Recognising panoramas. In *Proceedings of ICCV*, page 1218, 2003.

[31] H. Cai and J. Y. Zheng. Key views for visualizing large spaces. *Journal of Visual Communication and Image Representation*, 20(6):420–427, 2009.

[32] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices. In *Proceedings of CVPR*, pages 737–744. IEEE Computer Society, 2011.

[33] H. F. Chen and P. Meer. Robust regression with projection based m-estimators. In *Proceedings of ICCV*, pages 878–885, 2003.

[34] T. Chin, H. Wang, and D. Suter. Robust fitting of multiple structures: The statistical learning approach. In *Proceedings of ICCV*, pages 413–420. IEEE Computer Society, 2009.

[35] S. Choi, T. Kim, and W. Yu. Performance evaluation of RANSAC family. In *Proceedings of BMVC*, pages 1–12. BMVA, 2009.

[36] H. Chui and A. Rangarajan. A new algorithm for non-rigid point matching. In *Proceedings of CVPR*, pages II: 44–51, 2000.

[37] O. Chum and J. Matas. Matching with PROSAC - PROgressive SAmple Consensus. In *Proceedings of CVPR*, volume 1, pages 220–226, 2005.

[38] O. Chum and J. Matas. Optimal Randomized RANSAC. *IEEE Transactions on PAMI*, 30(8):1472–1482, August 2008.

[39] O. Chum, J. Matas, and J. Kittler. Locally optimized RANSAC. In *Proceedings of DAGM*, pages 236–243, 2003.

[40] O. Chum, M. Perdoch, and J. Matas. Geometric min-hashing: Finding a (thick) needle in a haystack. In *Proceedings of CVPR*, pages 17–24. IEEE Computer Society, 2009.

[41] O. Chum, T. Werner, and J. Matas. Two-view geometry estimation unaffected by a dominant plane. In *Proceedings of CVPR*, pages I: 772–779, 2005.

[42] D. Crandall, A. Owens, N. Snavely, and D. P. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *Proceedings of CVPR*, pages 3001–3008. IEEE Computer Society, 2011.

[43] F. Dellaert, S. M. Seitz, C. E. Thorpe, and S. Thrun. Structure from motion without correspondence. In *Proceedings of CVPR*, volume 2, pages 557–564, 2000.

[44] J. Domke and Y. Aloimonos. A probabilistic framework for correspondence and egomotion. In *Proceedings of WDV06*, pages 232–242, 2006.

[45] M. Douze, H. Jegou, H. Sandhawalia, L. Amsaleg, and C. Schmid. Evaluation of gist descriptors for web-scale image search. In *CIVR*. ACM, 2009.

[46] Z. T. Feng Han and S. C. Zhu. Range image segmentation by an effective jump-diffusion method. *IEEE Transactions on PAMI*, 26(9):1138–1153, 2004.

[47] M. A. Fishler and R. C. Boles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated pages cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[48] A. W. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *Proceedings of CVPR*, volume 1, page 125, 2001.

[49] J. Frahm and M. Pollefeys. RANSAC for (quasi-) degenerate data (QDEGSAC). In *Proceedings of CVPR*, volume 1, pages 453–460. IEEE Computer Society, 2006.

[50] J.-M. Frahm, P. F. Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, and S. Lazebnik. Building rome on a cloudless day. In *Proceedings of ECCV*, pages 368–381. Springer-Verlag, 2010.

[51] J. M. Frahm and M. Pollefeys. RANSAC for (Quasi-)Degenerate data (QDEGSAC). In *Proceedings of CVPR*, pages I: 453–460, 2006.

[52] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. In *Proceedings of CVPR*, pages 1434–1441. IEEE Computer Society, 2010.

[53] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on PAMI*, 32(8):1362–1376, August 2010.

[54] D. Gamerman. *Markov Chain Monte Carlo*. Chapman and Hall, 1997.

[55] B. Georgescu and P. Meer. Balanced recovery of 3d structure and camera motion from uncalibrated image sequences. In *Proceedings of ECCV*, page II: 294 ff., 2002.

[56] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. Seitz. Multi-view stereo for community photo collections. In *Proceedings of ICCV*, pages 1–8. IEEE Computer Society, 2007.

[57] L. Goshen and I. Shimshoni. Balanced exploration and exploitation model search for efficient epipolar geometry estimation. *IEEE Transactions on PAMI*, 30(7):1230–1242, 2008.

[58] L. Goshen and I. Shimshoni. Guided sampling via weak motion models and outlier sample generation for epipolar geometry estimation. *International Journal of Computer Vision*, 80(2):275–288, 2008.

[59] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[60] U. Grenander and M. Miller. Representations of knowledge in complex systems. *Journal of Royal Statistical Society, Series B*, 56(4):549–603, 1994.

[61] A. Hakeem, R. Vezzani, M. Shah, and R. Cucchiara. Estimating geospatial trajectory of a moving camera. In *Proceedings of ICPR*, volume 2, pages 82 –87, Hong Kong, 2006. IAPR.

[62] R. M. Haralick, C. N. Lee, X. Zhuang, V. G. Vaidya, and M. B. Kim. Pose estimation from corresponding point data. In *Proceedings of CVWS87*, pages 258–263, 1987.

[63] C. Harris and M. Stephens. A combined corner and edge detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[64] R. Hartley. In defense of the 8-point algorithm. *IEEE Transactions on PAMI*, 19(6):580–593, 1997.

[65] R. I. Hartley. In defense of the 8-point algorithm. *IEEE Transactions on PAMI*, 19(6):580–593, 1997.

[66] R. I. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, New York, NY, USA, 2000.

[67] T. Hassan, C. Ellum, and N. El-Sheimy. Bridging land-based mobile mapping using photogrammetric adjustments. In *ISPRS Commission I Symposium. From Sensors to Imagery*, Marne-laVallée, FRANCE, July 2006.

[68] K. Heath, N. Gelfand, M. Ovsjanikov, M. Aanjaneya, and L. Guibas. Image webs: Computing and exploiting connectivity in image collections. In *Proceedings of CVPR*, pages 3432 –3439. IEEE Computer Society, 2010.

[69] T. S. Huang and A. N. Netravali. Motion and structure from feature correspondences: A review. *Proceedings of the IEEE*, 82(2):252–268, February 1994.

[70] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of ECCV*, pages 304–317. Springer-Verlag, 2008.

[71] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *Proceedings of CVPR*, pages 506–513. IEEE Computer Society, 2004.

[72] D. Keren. A probabilistic method for point matching in the presence of noise and degeneracy. *Journal of Mathematical Imaging and Vision*, 33(3):338–346, 2009.

[73] S. M. Khan, F. Rafi, and M. Shah. Where was the picture taken: Image localization in route panoramas using epipolar geometry. In *Proceedings of ICME*, pages 249–252. IEEE Computer Society, 2006.

[74] G. Li and Y. Tsin. Globally optimal affine epipolar geometry from apparent contours. In *Proceedings of ICCV*, pages 96–103. IEEE Computer Society, 2009.

[75] X. Li, C. Wu, C. Zach, S. Lazebnik, and J. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *Proceedings of ECCV*, pages 427–440. Springer-Verlag, 2008.

[76] Longuet-Higgins. A computer algorithm for reconstructing from two projections. *Nature*, 293:133–135, 1981.

[77] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[78] Q. T. Luong and O. D. Faugeras. The fundamental matrix: Theory, algorithms, and stability analysis. *International Journal of Computer Vision*, 17(1):43–75, January 1996.

[79] J. Maciel and J. Costeira. A global solution to sparse correspondence problems. *IEEE Transactions on PAMI*, 25(2):187–199, February 2003.

[80] A. Makadia, C. Geyer, S. Sastry, and K. Daniilidis. Radon-based structure from motion without correspondences. In *Proceedings of CVPR*, pages I: 796–803, 2005.

[81] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[82] J. Matas, S. Obdrzalek, and O. Chum. Local affine frames for wide-baseline stereo. In *Proceedings of ICPR*, pages 363–366, 2002.

[83] P. Meer, D. Mintz, A. Rosenfeld, and D. Kim. Robust regression methods for computer vision: A review. *International Journal of Computer Vision*, 6(1):59–70, 1991.

[84] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on PAMI*, 27(10):1615–1630, 2005.

[85] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. J. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72, November 2005.

[86] L. Moisan and B. Stival. A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *International Journal of Computer Vision*, 57(3):201–218, 2004.

[87] A. I. Mourikis, N. Trawny, S. I. Roumeliotis, A. E. Johnson, A. Ansar, and L. Matthies. Vision-aided inertial navigation for spacecraft entry, descent, and landing. *IEEE Transactions on Robotics*, 25(2):264–280, 2009.

[88] D. R. Myatt, P. H. S. Torr, S. J. Nasuto, J. M. Bishop, and R. Craddock. NAPSAC: High noise, high dimensional robust estimation - it's in the bag. In *Proceedings of BMVC*, volume 2, pages 458–467, 2002.

[89] K. Ni, H. Jin, and F. Dellaert. GroupSAC: Efficient consensus in the presence of groupings. In *Proceedings of ICCV*, pages 2193–2200. IEEE Computer Society, 2009.

[90] D. Nistér. Preemptive RANSAC for Live Structure and Motion Estimation. In *Proceedings of ICCV*, page 199, 2003.

[91] D. Nister. An efficient solution to the five-point relative pose problem. *IEEE Transactions on PAMI*, 26(6):756–777, June 2004.

[92] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proceedings of CVPR*, pages 2161–2168. IEEE Computer Society, 2006.

[93] J. Oliensis. A critique of structure-from-motion algorithms. *Computer Vision and Image Understanding*, 84(3):407–408, December 2001.

[94] J. Oliensis and Y. Genc. Fast and accurate algorithms for projective multi-image structure from motion. *IEEE Transactions on PAMI*, 23(6):546–559, June 2001.

[95] A. Oliva and A. Torralba. Modeling the shape of the scene : A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[96] M. Perdoch, J. Matas, and O. Chum. Epipolar geometry from two correspondences. In *Proceedings of ICPR*, pages 215–219. IEEE Computer Society, 2006.

[97] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proceedings of ICCV*, pages 316–336. IEEE Computer Society, 2008.

[98] M. Pollefeys, F. Verbiest, and L. J. V. Gool. Surviving dominant planes in uncalibrated structure and motion recovery. In *Proceedings of ECCV*, pages 837–851, 2002.

[99] R. Raguram, J. Frahm, and M. Pollefeys. A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus. In *Proceedings of ECCV*, pages 500–513. Springer, 2008.

[100] R. Raguram, J. Frahm, and M. Pollefeys. Exploiting uncertainty in random sample consensus. In *Proceeding of ICCV*, pages 2074–2081. IEEE Computer Society, 2009.

[101] R. Roncella. Photogrammetric bridging of GPS outages in mobile mapping. *Proceedings of SPIE*, 5665:308–319, 2005.

[102] D. J. Rosenkrantz, R. E. Stearns, and P. M. L. II. An analysis of several heuristics for the traveling salesman problem. *SIAM Journal on Computing*, 6(3):563–581, 1977.

[103] S. Rozenfeld and I. Shimshoni. The Modified pbM-Estimator Method and a Runtime Analysis Technique for the RANSAC Family. In *Proceedings of CVPR*, pages I: 1113–1120, 2005.

[104] S. Santini and R. Jain. Similarity measures. *IEEE Transactions on PAMI*, 21(9):871–883, 1999.

[105] S. Saripalli, J. F. Montgomery, and G. S. Sukhatme. Vision-based autonomous landing of an unmanned aerial vehicle. In *Proceedings of ICRA*, pages 2799–2804. IEEE Computer Society, 2002.

[106] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or 'how do i organize my holiday snaps?'. In *Proceedings of ECCV*, pages 414–431. Springer-Verlag, 2002.

[107] G. Scott and H. Longuet Higgins. An algorithm for associating the features of two images. *RoyalP*, B-244:21–26, 1991.

[108] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of CVPR*, pages I: 519–528. IEEE Computer Society, 2006.

[109] L. S. Shapiro and J. M. Brady. Feature-based correspondence: An eigenvector approach. *Image and Vision Computing*, 10(5):283–288, June 1992.

[110] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In *Proceedings of CVPR*, pages 3013–3020. IEEE Computer Society, 2012.

[111] E. H. Shumate. The Radius of Curvature in the Prime Vertical. *International Test and Evaluation Association Journal*, pages 159–163, 2009.

[112] B. Sinopoli, M. Micheli, G. Donato, and T. J. Koo. Vision based navigation for an unmanned aerial vehicle. In *Proceedings of ICRA*, pages 1757–1764. IEEE Computer Society, 2001.

[113] G. Slabaugh. Computing Euler angles from a rotation matrix. `http://www.gregslabaugh.name/publications/euler.pdf`.

[114] N. Snavely. *Bundler: Structure from Motion for Unordered Image Collections*. Online, May 2009.

[115] N. Snavely, S. Seitz, and R. Szeliski. Skeletal graphs for efficient structure from motion. In *Proceedings of CVPR*, pages 1–8. IEEE Computer Society, 2008.

[116] C. V. Stewart. MINPRAN: A new robust operator for computer vision. *IEEE Transactions on PAMI*, 17(10):925–938, 1995.

[117] H. Stewénius, D. Nistér, F. Kahl, and F. Schaffalitzky. A minimal solution for relative pose with unknown focal length. *Image and Vision Computing*, 26(7):871–877, 2008.

[118] L. Tierney. Markov chains for exploring posterior distributions. *Annal of Statistics*, 22:1701–1728, 1994.

[119] W. S. Tong, C. K. Tang, and G. Medioni. Simultaneous two-view epipolar geometry estimation and motion segmentation by 4d tensor voting. *IEEE Transactions on PAMI*, 26(9):1167–1184, September 2004.

[120] B. J. Tordoff. Guided-MLESAC: Faster image transform estimation by using matching priors. *IEEE Transactions on PAMI*, 27(10):1523–1535, 2005. Member-David W. Murray.

[121] P. Torr. A Structure and Motion Toolkit in Matlab. `http://www.mathworks.com/matlabcentral/fileexchange/4576`, March 2004.

[122] P. Torr and D. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *International Journal of Computer Vision*, 24(3):271–300, 1997.

[123] P. H. S. Torr. *Outlier detection and Motion Segmentation, PhD Thesis*. Dept. of Engineering Science, University of Oxford, 1995.

[124] P. H. S. Torr. An assessment of information criteria for motion model selection. In *Proceedings of CVPR*, page 47, 1997.

[125] P. H. S. Torr. Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *International Journal of Computer Vision*, 50(1):35–61, 2002.

[126] P. H. S. Torr and C. Davidson. IMPSAC: Synthesis of Importance Sampling and Random Sample Consensus. *IEEE Transactions on PAMI*, 25(3):354–364, 2003.

[127] P. H. S. Torr, A. W. Fitzgibbon, and A. Zisserman. The problem of degeneracy in structure and motion recovery from uncalibrated image sequences. *International Journal of Computer Vision*, 32(1):27–44, 1999.

[128] B. Triggs. Joint feature distributions for image correspondence. In *Proceedings of ICCV*, volume 2, pages 201–208. IEEE Computer Society, 2001.

[129] P. H. S. Torr and A. Zisserman. MLESAC: a new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78(1):138–156, 2000.

[130] P. H. S. Torr, A. Zisserman, and S. J. Maybank. Robust detection of degenerate configurations while estimating the fundamental matrix. *Computer Vision and Image Understanding*, 71(3):312–333, 1998.

[131] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – a modern synthesis. In *Vision Algorithms: Theory and Practice*, volume 1883 of *Lecture Notes in Computer Science*, pages 298–372, 2000.

[132] A. Welsh. On M-processes and M-estimation. *Annal of Statistics*, 17(1):337–361, 1989.

[133] A. Wu, E. Johnson, and A. Proctor. Vision-aided inertial navigation for flight control. *Journal of Aerospace Computing, Information and Communication*, pages 348–360, 2005.

[134] K.-H. Yap, T. Chen, Z. Li, and K. Wu. A comparative study of mobile-based landmark recognition techniques. *IEEE Intelligent Systems*, 25(1):48–57, 2010.

[135] W. Zhang and J. Kosecka. Image based localization in urban environments. In *Proceedings of 3DPVT*, pages 33–40, Washington DC, USA, 2006. IEEE Computer Society.

[136] Z. Zhang. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision*, 27(2):161–195, 1998.

[137] Y. Zheng, S. Sugimoto, and M. Okutomi. A branch and contract algorithm for globally optimal fundamental matrix estimation. In *Proceedings of CVPR*, pages 2953–2960. IEEE Computer Society, 2011.

[138] Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the world: Building a web-scale landmark recognition engine. In *Proceedings of CVPR*, pages 1085–1092. IEEE Computer Society, 2009.

[139] W. Zhou, H. Li, Y. Lu, and Q. Tian. Large scale image search with geometric coding. In *Proceedings of ACM Multimedia*, pages 1349–1352. ACM, 2011.

[140] Y. Zhou, H. Leung, and M. Blanchette. Sensor alignment with Earth-centered Earth-fixed (ECEF) coordinate system. *IEEE Transactions on AES*, 35(2):410 –418, Apr. 1999.

Appendices

Appendix A Glossary of Terms

Wide-area sparse-view datasets : Collection of images of a geographically wide-spread region such that there are very few pairs of images that match, or in other words, the views are sparse.

Wide baseline : The distance between the cameras used to take a pair of images.

Random Sampling : Randomly selecting a data point.

Model : A set of data points and its parameters.

Sample : Set of data points required to produce a model.

Degenerate sample : A sample of data points insufficient to parameterize a model completely because of redundant information in the data points in the sample.

Correspondel : Sample of (7 or 8) correspondences required to produce a epipolar geometry model.

Transitive closure : If A is connected to B and B is connected to C, check whether A is connected to C.

Point features : Point features in any image are the detected corners in it. Corners are the pixels in an image that are distinguishable from its neighborhood pixels.

Photometry : Properties of an image due to its pixel values or appearance is referred to as photometry of an image.

Geometry : Properties of an image due to locations of pixels in it is referred to as geometric properties.

SIFT features : Scale Invariant Feature Transform (SIFT) are point features that are likely to get detected and described similarly even when the size of the image is changed.

Correspondences: In a pair of images, one-to-one matching point features from one image to the other are called correspondences.

Putative correspondences: Set of correspondences that are tentative and might not all be correct are called putative correspondences.

Inlier correspondences: Inlier correspondences are the correct correspondences in a set of putative correspondences.

Outlier correspondences: Outlier correspondences are the incorrect correspondences in a set of putative correspondences.

Inlier rate: Rate of correct correspondences in a set of putative correspondences.

Outlier rate: Rate of incorrect correspondences in a set of putative correspondences.

Image match scores or image similarity score: Degree of similarity between a pair of images ranging from 0 to 1 such that 0 is the match between two non-matching images and 1 is the value of maximally matching images.

Image dissimilarity score: Degree of dissimilarity between a pair of images ranging from 0 to 1. The sum of similarity score and dissimilarity score is 1.

Epipolar geometry: Epipolar geometry or camera geometry is the strongest constraint on corresponding points between a pair of images. Corresponding points between a pair of images must triangulate to a 3D point and this triangulation should define the orientation of one camera with respect to the other. Epipolar geometry is captured in the form of a matrix called the fundamental matrix. Fundamental matrix is also referred as extrinsic calibration parameters.

Extrinsic camera calibration parameters: The camera calibration parameters between a pair of cameras, that is, position and orientation of one camera with respect to the other camera.

Intrinsic camera calibration parameters: The camera calibration parameters of a single camera, that is, the focal length of the camera, pixel aspect ratio, skew and optical center of the camera.

Photometric match scores: Match score estimated using photometrically established putative correspondences.

Geometric match scores: Match score estimated using geometrically determined inlier correspondences.

Spanning tree: A spanning tree of a connected graph is a sub-graph connecting all nodes with minimum number of edges. A spanning tree might not be unique.

Minimum Spanning Tree: A Minimum Spanning Tree of a connected graph is a sub-graph connecting all nodes using minimum number of edges with the lowest sum of edge weights. A minimum spanning tree also might not be unique.

Degree 2 constrained Minimum Spanning Tree: A degree-2 constrained Minimum Spanning Tree of a connected graph is a sub-graph connecting as many nodes as possible such that no nodes in the graph has a degree greater than 2. A degree-2 constrained Minimum Spanning Tree is also a Minimum Hamiltonian Path. A connected graph might not necessarily have a connected degree-2 constrained Minimum Spanning Tree.

Basal graphs: The term basal means minimal and foundational. Basal graphs are spanning forests in a graph. They are minimal in having minimum number of edges and foundational for applications. We propose minimal spanning forests and minimum degree-2 constrained spanning forests

Global Positioning System (GPS): A satellite system sending position signals to a receiver device.

Magnetometer and Accelerometer: Sensors that sense the orientation of a device.

Appendix B Notations

Symbol	Represents
$\mathcal{V}$	Set of N images, $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_N\}$.
$\mathcal{E}$	Set of weighted edges between images.
$\mathcal{G}(\mathcal{V}, \mathcal{E} = \phi)$	Graph connecting images with geometrically weighted edges.
$\mathcal{G}'(\mathcal{V}, \mathcal{E} = \varphi)$	Graph connecting images with photometrically weighted edges.
ϕ_{ij}^1	Geometric weights (inlier rate) on $\mathcal{E}$.
ϕ_{ij}^2	Geometric weights (number of inliers) on $\mathcal{E}$.
φ_{ij}	Initial photometric weight of $\mathcal{E}$.
κ	Constant correlating φ_{ij} to geometric inlier rate.
${\varphi_{ij}}^t$	Updated photometric weight of $\mathcal{E}$ after t iteration of geometric verification.
n_i	Number of features in $\mathcal{I}_i$
c	Number of connected components Λ in $\mathcal{V}$
s_i	Number of splits basal paths in Λ
$\boldsymbol{\Upsilon}$	Basal Tree Graphs of $\mathcal{G}'$.
$\boldsymbol{\Upsilon}^{\mathbf{a}}$	Basal Tree Graphs of a times updated $\mathcal{G}'$.
$\boldsymbol{\Upsilon}^*$	Basal Tree Graphs of a times updated $\mathcal{G}'$.
$\boldsymbol{\Lambda}$	Expanded Basal Graphs of $\mathcal{G}'$.
$\boldsymbol{\Pi}$	Basal Path Graphs.
$\mathbf{S}_{ij}$	Matrix of feature similarity between features of images $\mathcal{I}_i$ and $\mathcal{I}_j$.
$\mathbf{X}_{ij}$	Set of putative correspondences between $\mathcal{I}_i$ and $\mathcal{I}_j$.
$\lvert\mathbf{X}_{ij}\rvert$	Number of putative correspondences between $\mathcal{I}_i$ and $\mathcal{I}_j$.
ρ_{ij}^k	Similarity value in S_{ij} corresponding to kth putative correspondence in X_{ij}.
$\rho_{ij}^{k_r}$	Second highest similarity value along the row of S_{ij} corresponding to kth putative correspondence in X_{ij}.
$\rho_{ij}^{k_c}$	Second highest similarity value along the column of S_{ij} corresponding to kth putative correspondence in X_{ij}.
x_{ij}^k	kth putative correspondence between $\mathcal{I}_i$ and $\mathcal{I}_j$.
w_{ij}^k	Photometric confidence weight on kth putative correspondence between $\mathcal{I}_i$ and $\mathcal{I}_j$.
s	Size of minimal sample.
d	A vector of size s containing indices of sampled correspondences.
θ^t	Minimal sample set of correspondences chosen from $\mathbf{X}_{ij}$ in the tth iteration.
θ^*	Optimal minimal sample set of correspondences chosen from $\mathbf{X}_{ij}$.
$\mathbf{V}_{ij}$	Scatter Matrix for θ
$\mathbf{W}_{ij}$	Information Tensor for θ

Appendix B (continued)

Symbol	Represents
$\mathbf{F}_{ij}$	Fundamental matrix between $\mathcal{I}_i$ and $\mathcal{I}_j$.
δ_{ij}^k	Sampson's Distance of kth putative correspondence from F_{ij}.
$\mu(\mathbf{F}_{ij}(\theta)\|\mathbf{X}_{ij})$	Geometric M-estimate of the number of inliers in X_{ij} using θ_{ij}.
ω_{ij}	0/1 degeneracy indicator for θ_{ij}.
μ_{ij}	Geometric M-estimate of the number of inliers in X_{ij} using ${\theta_{ij}}^*$.
Γ^1	Threshold on the number of inliers.
Γ^2	Lower threshold on inlier rate.
Γ^3	Upper threshold on inlier rate.
$\mathbf{K}_{ij}$	Intrinsic calibration matrix between $\mathcal{I}_i$ and $\mathcal{I}_j$.
$\mathbf{P}_i^E$	Geodetic lat, long and altitude (l_i, g_i, h_i) coordinates in meters.
$\mathbf{P}_i$	ECEF coordinates (x_i, y_i, z_i) in meters.
$\mathbf{O}_i$	Camera orientation angles $(\psi_i^1, \psi_i^2, \psi_i^3)$
$r_\ominus$	Radius of Earth in meters
$e_\ominus$	Eccentricity of Earth in meters
$\mathbf{Q}_{ij}^v$	Vision based unit quaternion rotation vectors between $\mathcal{I}_i$ and $\mathcal{I}_j$.
$\mathbf{Q}_{ij}^m$	Magnetometer based unit quaternion rotation vectors between $\mathcal{I}_i$ and $\mathcal{I}_j$.
T_{ij}^v	Vision based unit translation vector between $\mathcal{I}_i$ and $\mathcal{I}_j$.
$\mathbf{T}_{ij}^g$	GPS based unit translation vector between $\mathcal{I}_i$ and $\mathcal{I}_j$.
$\mathbf{D}_{ij}^G$	Dot product of vision and GPS based unit translation vectors between cameras for $\mathcal{I}_i$ and $\mathcal{I}_j$.
$\mathbf{D}_{ij}^M$	Dot product of vision and GPS based unit rotation vector between cameras for $\mathcal{I}_i$ and $\mathcal{I}_j$.
$\mathbf{C}_i^G$	Confidence on GPS tag in $\mathcal{I}_i$.
$\mathbf{C}_i^M$	Confidence on magnetometer tag in $\mathcal{I}_i$.
$\mathbf{CC}_i^G$	Color code to show errors in GPS tag in $\mathcal{I}_i$.
$\mathbf{CC}_i^M$	Color code to show errors in magnetometer tag in $\mathcal{I}_i$.
$\mathbf{R}(\psi_{ij}^{m_z})$	Magnetometer roll.
$\mathbf{R}(\psi_{ij}^{m_y})$	Magnetometer pitch.
$\mathbf{R}(\psi_{ij}^{m_x})$	Magnetometer yaw.
$\mathbf{PR}_{ij}'$	Projection matrix associated with images $\mathcal{I}_i$ and $\mathcal{I}_j$.

Appendix C Significance Test for Simple Linear Regression

Given two measurements $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_n\}$, we want to find a regression line $Y = mX + C$. Let X be the independent variable and Y be the dependent variable and let C be a constant.

$$\arg\min_m \sqrt{(\sum (Y - mX - C).(Y - mX - C))} \tag{C.1}$$

On solving the above equation by equating the derivative to zero, m can be obtained as

$$m = \frac{\overline{X.Y} - \overline{X}.\overline{Y}}{\overline{X^2} - \overline{X}^2} \tag{C.2}$$

$$m = r\frac{\sigma_y}{\sigma_x} \tag{C.3}$$

where r is the correlation coefficient given by

$$r = \frac{\sum (x_i - \overline{X}) \sum (y_i - \overline{Y})}{\sqrt{\sum (x_i - \overline{X})^2 \sum (y_i - \overline{Y})^2}} \tag{C.4}$$

Let the null hypothesis be : There exists no straight line relationship between X and Y. The correlation coefficient r is used to test the null hypothesis. The relationship is small for the correlation coefficient values less that 0.3, medium for the correlation coefficient values greater than 0.3 and less than 0.5, large for the correlation coefficient values greater than 0.5.

Appendix D Linear Regression with Zero Intercept

Given two measurements $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_n\}$, we want to find a regression line passing through the origin. Let X be the independent variable and Y be the dependent variable.

$$\arg\min_m \sqrt{((Y - mX).(Y - mX))} \tag{D.1}$$

On solving the above equation by equating the derivative to zero, m can be obtained as

$$m = \frac{\sum(X.Y)}{\sum X.X} \tag{D.2}$$

However, the correlation coefficient cannot be determined using the zero intercept model.

About the Author

Aveek Shankar Brahmachari received the B.Tech. degree in Computer Science and Engineering from Indian School of Mines, Dhanbad, India in 2004. Thereafter, he worked as a scientist with Image Analysis Group, DEAL, DRDO in India for 3 years. He received his MS degree in Computer Science and Engineering from University of South Florida, Tampa in 2009. He is currently a PhD candidate at the University of South Florida. His research interests include image processing, computer vision, pattern recognition and machine learning. He is a student member of the IEEE.